FUNNELBACK AND ME

Celebrating Thirty Years of Funnelback Technology
1991 — 2021

DAVID HAWKING

Funnelback and Me

Celebrating Thirty Years of Funnelback Technology
1991 – 2021

"Search is life!"

David Hawking
david.hawking@acm.org

ISBN 978-0-6451743-5-9 PDF; 978-0-6451743-6-6 Print

Typeset in 11pt Palatino using LaTeX. Last edited: 14 May 2022
Cover design: Jack Griffiths-Hawking
Back cover photos taken by helpful colleagues using the author's and/or Stuart Beil's cameras.

In 1991, while Head Programmer in the Department of Computer Science at ANU, I wrote a text search program in order to learn how to program a parallel supercomputer. That led to research in the field of Information Retrieval and I gradually migrated into a research role in the Advanced Computational Systems (ACSys) CRC, obtaining a PhD by published work. In 1998 I took up a research scientist position in CSIRO Mathematical and Information Sciences, while remaining a member of ACSys. Working with ACSys colleagues, we extended my text retrieval system into an intranet search engine (P@NOPTIC) which launched on `anu.edu.au` in 1999. As a CSIRO scientist I was required to generate substantial external earnings. I chose to do this by creating a virtuous cycle – licensing P@NOPTIC to customers, learning about their real-world search problems, researching solutions, improving the product, and licensing to more customers. This was slow to build up but eventually quadrupled the external earnings target. Indeed, the P@NOPTIC cottage industry became too big to remain in CSIRO and, in 2005, the technology was spun off as Funnelback Pty Ltd. CSIRO exited in 2009 when it sold Funnelback to Squiz Pty Ltd. By 2020, Funnelback had earned tens of millions of dollars, employed a peak of over 50 people in Australia, the UK and the US, and, for many organisations, dramatically improved the ability of stakeholders to find internal information.

I'm quite proud of what was achieved, but in this account I try to analyse why Google rather than Funnelback took over the world of search. The 30 year history of Funnelback and its predecessor research is full of colourful characters, weird and wonderful projects, bizarre customer behaviour, the dejection early on of being declared a failure, the pain of losing out on deals, the thrill of closing them, the excitement of winning awards, and the satisfaction of solving customer problems.

Contents

Acknowledgements

This is essentially a personal account of the history of Funnelback Pty Ltd and its predecessors: P@NOPTIC and S@NITY; and my team's research at ANU, the ACSys Cooperative Research Centre, and CSIRO. Fortunately, many of the gaps and biases in my personal recollections have been filled by contributions from my former colleagues, and from some of the people who worked at Funnelback after my departure.

I gratefully acknowledge contributions from: Peter Bailey, Stuart Beil, Narelle Bortolin, Luke Butters, Michael Carter, Prathima Chandra, Nick Craswell, Francis Crimmins, Alwyn Davis, George Ferizis, Gordon Grace, Natalie Grech, Jack Griffiths-Hawking, Nicolas Guillaumin, Mandhakini Iyer, Tim Jones, Brett Matson, Will Parkinson, Sonia Piton, Ben Pottier, Annie Pritchard, Tom Rowlands, Matt Sheppard, Matt Taylor, Peter Thew, Paul Thomas, Ben Tilley, Trystan Upstill, Phil Widdop, Shaw Xiao.

I also wish to express my gratitude to:

- The late Paul Thistlewaite for being a great collaborator, for ensuring that information retrieval was an ACSys research area and for mapping out the AWESOME vision which eventually led to S@NITY, P@NOPTIC, and Funnelback.
- Heads of ANU Computer Science, Robin Stanton, Brian Molinari and Chris Johnson, for permitting and facilitating my transition from system administrator to researcher.
- ACSys directors, Robin Stanton, Michael McRobbie, John O'Callaghan, and Darrell Williamson for supporting the PASTIME, TAR, and WAR projects.
- CSIRO managers: Ross Wilkinson, Rhys Francis, Murray Cameron, and Alex Zelinsky for recruiting me and supporting what became a very successful commercialisation of research.
- Anthea Roberts and Sarah Savage for providing exceptional HR support within CSIRO.
- Funnelback leaders over the years for giving me the freedom to pursue my ideas and to continue to be an active member of the scientific research community. I hope I justified your faith in me.

Chapter 1

Introduction

The period of my involvement in Funnelback, and the research and development which led up to it – from early 1991 to late 2013 – was wildly exciting. I learned many new skills; I worked harder than at any other time in my life; I stretched myself way beyond my comfort zone and did things I never expected I would be able to do; I interacted with impressive people who radically changed the world; and I was a privileged ring-side observer of the rise to dominance of web search engines like Google. There were times of huge anticipation, wild excitement, and deep disappointment.

I had no prior inkling that I would enjoy being involved in commerce but I found myself working very hard to close deals and felt an enormous thrill when I occasionally closed one. People's jobs depended upon sales!

My career as a scientific researcher was also unexpected. In 1991 I was working as a system and network administrator, having long abandoned a PhD in computer animation and any thought of an academic career. Later, thanks to my work on what became Funnelback technologies, I gained a PhD, an honorary doctorate and a reasonably successful research track record. I obtained great satisfaction from doing good science, and from whole-greater-than-sum-of-parts interactions with colleagues and PhD students.

I take enormous pride in the fact that our research led to the creation of a company which improved the quality of search at hundreds of organisations, earned many tens of millions of dollars in revenue, attained a peak of around 50 high-tech jobs, and significantly improved people's ability to find information. Tiny compared to internet behemoths but nonetheless satisfying.

None of this would have happened without the vital contributions of the many scientific, technical and business colleagues whose roles are described in what follows. I apologize to current or former employees whose contributions I haven't fully recognized, due to inability to contact, lack of awareness, or failures of memory. The period after I left Funnelback in 2013 is, of necessity, covered in less detail.

1.1 Overview of the Funnelback Story

PADRE, the central component of Funnelback search technology, originated at ANU in 1991 as a spare-time project designed to help me learn to write programs for a massively parallel supercomputer. It morphed into a research project within the Advanced Computational Systems Cooperative Research Centre (ACSys CRC), and led to me gaining a PhD (by published work) in 1998. That year I joined CSIRO Mathematical and Information Sciences (CMIS) but continued to work part-time in ACSys. Commercialisation objectives within both ACSys and CSIRO led to the addition of other components of an enterprise search system and the launch of the first customer site in July 1999 – the Australian National University. At the time the system was called S@NITY, but, to avoid trademark issues, the name was soon changed to P@NOPTIC. Other customers followed and we were soon running a small business within CMIS. By 2005, we had achieved annual revenues in excess of a million dollars, causing potential problems for CSIRO:

- It's inappropriate for CSIRO, as a government agency, to compete with the private sector;
- Running a business requires spending a lot of time on things which cannot be classified as scientific research.

Accordingly, at the very end of 2005, the technology and some of the people were spun off as Funnelback Pty Ltd. I stayed as a CSIRO employee, but was seconded to Funnelback at about 50% of full-time. In 2008, I joined Funnelback as full-time Chief Scientist. In 2009, Funnelback was acquired by Squiz Pty Ltd, using vendor finance. Initially Funnelback operated largely independently of Squiz, but over time the two became more closely integrated. In late 2013 I left Funnelback, joining Microsoft for the opportunity to find out how a full-scale web search engine (Bing) works, and hopefully contribute to its success. Eventually, after the Funnelback CEO, Brett Matson, left in 2020, Funnelback was fully integrated into Squiz.

1.2 Definitions

An *information retrieval system*, sometimes known as a *search engine*, *text retrieval system*, or just *retrieval system* is a computer program associated with a *corpus* of documents,[1] whose job it is to receive queries from users and to return for each, a set or list of relevant or useful documents from the corpus. A *Boolean retrieval system* returns an unordered set of documents satisfying a logical query such as `(cat OR dog) NEAR (veterinarian OR vet)`. A *ranked retrieval system* returns a list of documents most likely to be useful to the searcher, ordered by descending order of probably utility. Retrieval systems are generally imperfect and usually retrieve some useless documents in the retrieved list (or set). For non-random systems, the further down a ranked list one goes, the more likely it is that a retrieved document is not useful.

In the world of information retrieval, it is traditional to assess whether a retrieved document is *relevant* – whether it relates to the topic of the information need. I prefer to think instead of *utility* – how useful a retrieved document is to the searcher. A relevant document may not be useful to the searcher – it may relate to the topic but fail to satisfy the searcher's need.

A *document* is the basic unit of retrieval, such as a newswire article, a web page, or an email message. It's definition is somewhat arbitrary – is a book a document, or do finer divisions such as chapters, sections or even paragraphs provide more useful units of retrieval?

A *query* is the text sent by a user to a text retrieval system in order to initiate retrieval. It is sometimes also known as a *request*. The simplest form of query is a *bag-of-words query*, consisting of a set of words, without operators, in which the order of the words is unimportant. For example, in a bag-of-words system, `australian national university` and `university national australian` would be expected to produce identical results.

Most retrieval systems support queries containing operators. Even ranked retrieval systems sometimes support logical operators such as AND, OR, and NOT, NEAR and FOLLOWED-BY. They often allow operators such as: double-quote, allowing the definition of phrases; $+$, meaning that a query term must be present in each retrieved document; and $-$, meaning that a query term must be absent in all retrieved documents. Some retrieval systems allow for left or right truncation, represented by an asterisk. For example `economic*` (right truncation) would match words starting with `economic` and `*alism` (left truncation) would match words ending with `alism`.

A query *term* is one element of a query. In a bag-of-words query, each word is a term. In an enhanced retrieval system, query terms may be compounds such as phrases, or proximity relations. In the following query:

```
+"david hawking" "australian national university" honorary
```

there are three top level terms: `+"david hawking"`, `"australian national university"` and `honorary`.

Many search engine customers refer to queries as *search terms*, but I find this confusing.

[1] A body or set of documents.

1.3 Reader's Guide

This book tells the story of Funnelback and its precursors in some detail. My goal is to be "entertaining and accurate" rather than "comprehensive." As noted in the Acknowledgements, the story is told from my own perspective, though I've interviewed and/or consulted many key players.

Large parts of the story are scientific or technical in nature. I have tried to write up these parts as clearly and simply as possible, but I understand if you want to skip them. I have tried to record the scientific questions we investigated, the software and datasets we built, and when we did those things. I certainly don't intend to claim scientific priority for everything we did. We were the first to do some things, but in many cases we were following up ideas due to others, or applying them in new areas.

Chapter 2 describes the development of the PADRE retrieval engine up to the point of the ANU S@NITY launch. Chapter 3 describes the development of the other system components needed to facilitate the launch of S@NITY at ANU. Chapter 4 describes search-related research conducted within CSIRO until I left CSIRO in 2008. Chapter 5 describes the commercialisation of P@NOPTIC within CSIRO. Chapter 6 recounts the first years of the Funnelback spin-off, prior to the takeover by Squiz. Chapter 7 recounts the years of Squiz ownership during which Funnelback operated mostly independently of Squiz. Chapter 8 brings us up to the present day, 2021, Chapter 9 records some frustrating experiences in the commercialisation story, and Chapter 10 asks "what if" questions. The last chapter records some overall impressions of Funnelback from people who have been part of the story.

It's possible that some readers will be interested in the story of the research underpinning or surrounding Funnelback while others will ignore it and be more interested in the human and commercial stories. It's not trivial to clearly delineate the two threads, but I have tried to mark the sections which make up the research story with ✳.

In the book, you will find a large number of web links to enable you to pursue topics in full detail. Obviously, I can't guarantee that these links will continue to be valid into the future. If you encounter a dead link, I can only suggest either searching for the topic, or looking for the original URL on the Wayback Machine (`https://archive.org/`).

Chapter 2

PADRE: 1991 to 1999

PADRE is the name for a small set of C programs which provide the document indexing and search capabilities of the Funnelback system. PADRE generally assumes that a corpus of documents already exists. Within Funnelback, PADRE relies on other system components to crawl or gather the document corpus, and to provide the user interface and other ancilliary services.

2.1 *1991: PADRE Origins

In early 1991 ANU operated two supercomputers, a Fujitsu VP100 vector processing system, and a massively parallel (16,384 node) Connection Machine CM2, and was about to acquire a third, a Fujitsu AP1000 cellular array processor with 128 nodes. Unlike the other two, the AP1000 would be located in the Computer Science Department (DCS) on the top floor of the Crawford Building (now the Beryl Rawson Building). As Head Programmer in the Department of Computer Science, I was to be the manager of the AP1000 and responsible for its installation and subsequent operation.

1991: Fujitsu marketing photo for the AP1000 featuring me (left) and Paul Mackerras. *Photo: Fujitsu Marketing*

Prior to delivery of the AP1000 I, along with Paul Mackerras and two Fujitsu Australia engineers, was sent to Kawasaki, Japan to learn how to provide hardware and software support on the first AP1000 outside Japan. I had already been trained as an Application Engineer on the CM2 and claim to be the only person to ever receive hardware training on both of those machines.

Some time after the AP1000 was successfully installed, an evangelist for the Thinking Machines Corporation came to ANU to stimulate usage of the CM2 and, no doubt, to promote further purchases. One of the CM2 applications he highlighted was the use of a CM2 by the Dow Jones Corporation to provide highly responsive search and retrieval over a vast collection of financial documents. I found it inspiring.

Since I felt that the manager of a parallel computer ought to know how to write parallel programs, I started to write a simple text retrieval application for the Fujitsu AP1000. A possible side benefit was that a retrieval system for the AP1000 would please Fujitsu[1] by expanding the range of applications for its new machine.

There seemed to be considerable scope to speed up text retrieval by exploiting parallelism. At the time, I was told that FreeWAIS, a text retrieval system associated with Brewster Kahle's Wide Area Information System, took about a week to index a gigabyte of text on a Sun workstation of the time. Once the text was loaded, the AP1000 could perform that task in a few minutes.

A contemporary project using a parallel computer (the ICL Distributed Array Processor (DAP)) for text retrieval was described in *High speed text retrieval from large databases on a massively parallel processor*[2] by Stewart Reddaway, who was in fact the principal architect of the DAP. I had the pleasure of conversing with Stewart when he visited CSIRO. Also in the UK, Andy MacFarlane of City University, London completed a PhD on parallel information retrieval in 2000, under the supervision of Stephen Robertson. I was able to arrange for funding for him to visit us at ACSys for three months in order to do his experimental work on our DEC Alpha Farm.

2.2 *Dictionary Searching

Coincidentally, an ANU IT person, Harriet Michell, transferred around this time to the Australian National Dictionary Centre (ANDC). ANDC was partly funded by Oxford University Press (OUP) to produce dictionaries of Australian English such as the Australian Concise Oxford Dictionary (ACOD), and the Australian Pocket Oxford Dictionary (APOD). OUP provided ANDC with electronic versions of several of the Oxford dictionaries including the 20-volume OED2, in SGML form.[3] ANDC also had access to a search and retrieval system called `PAT`, which had been developed at the University of Waterloo, and was based on the Patricia Tree data structure developed there by computer science researchers Gaston Gonnet, Ricardo Baeza-Yates, and Tim Snider.

Harriet approached DCS to see if there was any way computer science could provide assistance in improving on the tools they were using. One of the problems they were facing was ensuring consistency in the way things were written. For example, "air force" was sometimes written `air force` and sometimes `airforce`. ANDC reported that it wasn't easy with PAT to find all examples regardless of how they were written.

I volunteered to try to make a version of PAT which would run on the AP1000 and support more complicated search expressions. I was signed up as a member of ANDC and given access to the 550MB OED2 data set. In 1991, that seemed like a huge amount of data – at the time typical workstations were configured with up to 16MB of RAM and about 100MB of disk storage. The AP1000, when initially installed, had no disk storage but a total of 2GB of RAM (128×16MB). The plan was that the OED2 would be stored on the disks of the front-end machine, and when required, roughly equal segments would be loaded into the RAM of each AP1000 node. A master program running on the front-end would interact with the user, broadcasting queries to all nodes and aggregating an-

[1] Fujtsu had effectively donated the AP1000 to ANU in the hope that ANU would showcase the machine in the English speaking world, and increase its appeal by developing software for it.

[2] `https://www.sciencedirect.com/science/article/abs/pii/0306457391900862`

[3] SGML (Standard Generalised Markup Language) is a text markup language very similar to the later XML.

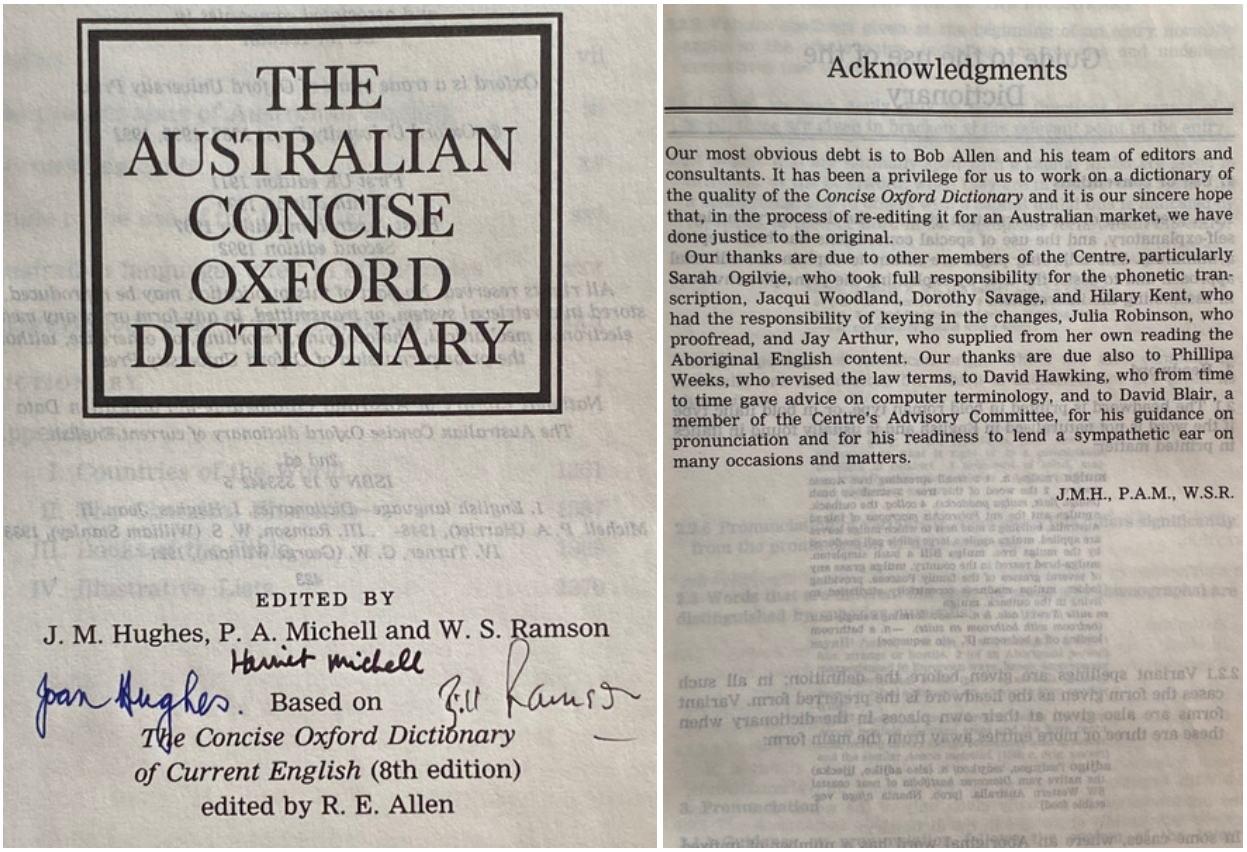
swers. Because each node would be responsible for less than 5MB of data, it was perfectly feasible to do brute force searches rather than relying on indexes, and to use sophisticated matching such as regular expressions.

The program I wrote was called `PADDY` – (Parallel Analysis of Dictionary Data – Yes!) PADDY succeeded to the extent of allowing more sophisticated search terms and achieving quite reasonable response times. However, it didn't implement some of the features supported by Patricia trees, such as being able to find the longest section of repeated text.

Unfortunately, using a program on the AP1000 was much less convenient than running PAT on your local workstation. Only one person could use the AP1000 at a time, so access needed to be scheduled, and although you could find answers in around 15 seconds, loading the data from the front end took something like 15 minutes.

Tragically, around this time Harriet was killed in a road accident near Nimmitabel, and opportunities for further technical cooperation with ANDC came to an end.

I wrote PADDY in my spare time, coding and debugging over breakfast. I wrote it in C, a language in which I was by no means expert, and remember receiving helpful suggestions from Robert Cohen, Paul Mackerras, and Andrew Tridgell.



THE AUSTRALIAN CONCISE OXFORD DICTIONARY

EDITED BY
J. M. Hughes, P. A. Michell and W. S. Ramson

Based on
The Concise Oxford Dictionary of Current English (8th edition)
edited by R. E. Allen

Acknowledgments

Our most obvious debt is to Bob Allen and his team of editors and consultants. It has been a privilege for us to work on a dictionary of the quality of the *Concise Oxford Dictionary* and it is our sincere hope that, in the process of re-editing it for an Australian market, we have done justice to the original.

Our thanks are due to other members of the Centre, particularly Sarah Ogilvie, who took full responsibility for the phonetic transcription, Jacqui Woodland, Dorothy Savage, and Hilary Kent, who had the responsibility of keying in the changes, Julia Robinson, who proofread, and Jay Arthur, who supplied from her own reading the Aboriginal English content. Our thanks are due also to Phillipa Weeks, who revised the law terms, to David Hawking, who from time to time gave advice on computer terminology, and to David Blair, a member of the Centre's Advisory Committee, for his guidance on pronunciation and for his readiness to lend a sympathetic ear on many occasions and matters.

J.M.H., P.A.M., W.S.R.

I'm proud of my small contribution to the Australian Concise Oxford Dictionary

I submitted a paper on PADDY[4] to the Australasian Supercomputing Conference in December 1991. By that stage I had enhanced the code to support the building of a word-start index, effectively a suffix array, and its use in simple searching. The paper reports that the index could be built in under three minutes. Searches for literal words were dramatically speeded up – 10 milliseconds to find all 327 occurrences of an example word. A paper I presented at the 1992 ANU-Fujitsu CAP Workshop,[5] reports further enhancements to PADDY and acknowledges the contributions of a vacation student Michael Hiron. It also speculates about the future of digital libraries based on such technology.

The PADDY query language provided primary operators such as word matching, phrase matching, prefix and suffix matching, and regular expression matching, which built sets of *matchpoints*. A

[4] `http://david-hawking.net/pubs/Hawking91.pdf`
[5] `http://david-hawking.net/pubs/Hawking92.pdf`

matchpoint is a reference to the character in the text where a match starts. For example if the query were `*alism` and the text were `A belief in American exceptionalism` then a matchpoint would be recorded for the 'e' at the start of `exceptionalism`.

Also provided were set operators such as UNION, INTERSECTION, NEAR, and FOLLOWED-BY which combined two sets of matchpoints into a single result set. The idea was to compose a query which produced a single result set, hopefully containing all and only the wanted answers. Like PAT, PADDY displayed the matches in the result set with a specified number of characters of pre and post context. There was no concept of "documents" and results were not ranked.

The paper included example regular expressions to find:

- All words ending in `ize`,
- All words containing `consider`,
- All words containing exactly three occurrences of the letter `i`, and
- All pallindromic headwords.

Apparently I ran a fourth-year Computer Science Honours course in this area in 1991. My memory of it has faded but Peter Bailey remembers:

> In semester 2, 1991 of the Computer Science Honours curriculum at ANU, Dave Hawking was the enthusiastic lecturer of our parallel programming course. As an Honours student, I'd started to interact more often with members of the department of Computer Science, so knew Dave a little bit as the Head Programmer who kept all our systems working. With the arrival of the Fujitsu AP1000, several of us doing Honours that year were already starting to be familiar with the challenges and opportunities of working with this state-of-the-art multicomputer.
>
> As a practical assignment, Dave combined his own emerging AP1000 text retrieval project of doing search over the OED with an exploration of different parallel sorting algorithms. We worked in groups of three, which was fun as well. The goal for each team was to implement one of a handful of different algorithms, measure its performance, and write a report describing the results. From memory, by chance my group was allocated the simplest possible approach – namely sorting results from Dave's PADDY program on each processor, then doing an N-way merge sort in the host controller. It turned out that this approach was overall both simple and efficient. Although the final merge sort was inherently serial (thus not making use of "smarter" parallel sorting algorithms), the orders-of-magnitude time differences in inter-processor communication relative to on-processor sorting meant that it had substantially faster performance. The only challenge that could arise was for very common terms, like `the`, where the sheer volume of results could overwhelm the memory limits of the host controller.
>
> Overall, the course taught me a lot about tradeoffs in memory access vs network access latencies, and that simple approaches are often worth trialling first. Dave's infectious enthusiasm for the problem space, both in terms of text search and parallel programming, and his gift for making the subject matter interesting and relevant, were a highlight among the courses that year. Searching the OED for all possible matches of some query in just a few seconds was very cool. I credit my lifelong professional interest in information retrieval to that course – and of course to Dave!

My very first scientific presentation

The first time I presented a paper at a scientific conference I was extremely nervous. Part-way through my talk, I became unable to see the audience or the room – the world suddenly went dark. I could hear a faint voice talking in the darkness and after a moment realised that it was mine. Reassured that I hadn't fainted, I continued and after a while reality returned.

The conference was the ANU-Fujitsu meeting held in the ANU Engineering Lecture Theatre in 1991. It was an amazing out-of-body experience and I never asked anyone whether they were aware of what I had experienced. No-one ever said anything.

After that I gave hundreds of presentations across dozens of countries, still occasionally feeling very nervous, but mostly enjoying the experience of interacting with an audience, and communicating knowledge. Interaction is the key – far better to have expressions of incomprehension or scepticism and react to them, than a totally unresponsive audience. I sometimes dreamed of what it must be like to engage a huge audience as Freddie Mercury and AC/DC did.

PADDY used the Boyer-Moore-Gosper (BMG) algorithm to allow faster search for literal strings. The straight-forward algorithm for finding occurrences of a search *pattern* such as

```
antidisestablishmentarianism
```

looks at every single character in the text data to see whether an occurrence of the pattern starts there. BMG achieves substantial speed-up by pre-computing a skip table for the pattern. The length of the antidisestablishmentarianism pattern is 28, so BMG first examines the 28th character of the text. If that character is a 'z', or any other character which doesn't occur in the pattern, the skip table tells it that it can skip forward a full 28 characters, since no match can occur before that. If the text character is an 'r' which occurs 7 characters from the end of the pattern, BMG can skip forward 6 places.

We'll come back to BMG later in the story.

In 1993, ANU's AP1000 was fitted with disk storage accessible from the nodes, and Peter Bailey joined me on the project. A 1993 paper we wrote for the Second Fujitsu Parallel Computing Workshop[6] details major steps toward a practical information retrieval system for the AP1000. In a nutshell, these improvements (implemented by Peter) allowed a user on a remote workstation to request the loading and querying of one of several text databases held on the AP1000 local disks.

How hard can it be to make a career in research?

[6] http://david-hawking.net/pubs/HawkingB93.pdf

2.3 ∗The Text Retrieval Conference

In 1992 or 1993 I was asked to make a poster about PADDY and present it at a showcase of exciting research within the yet-to-be-formally-constituted ANU Faculty of Engineering and IT. One of the visitors to the showcase was Richard Jones, then Chief Scientist at Computer Power Pty Ltd. Richard asked me if I was aware of TREC, the Text Retrieval Conference, organised by the National Institute of Standards and Technology (NIST) in Washington DC. He explained that it was a collaborative endeavour designed to improve the standard of text retrieval technology, and passed on the email address of Donna Harman, the TREC Co-ordinator.

You may not be surprised to learn that TREC was sponsored by US government agencies such as the National Security Agency (NSA) who (presumably) wanted faster and better means of extracting threat indications from huge volumes of transcripts from signals intercepts!

In the main evaluation task, *ad hoc retrieval*, NIST supplied TREC participants with a standard corpus of documents, and a set of 50 English language research topics. There were about 500,000 documents in the corpus, about 2GB in total. The corpus was made up of sub-corpora, each comprising documents from a single source. Content included US patents, news reports from Associated Press and the Wall Street Journal, Ziff-Davis IT magazine articles, plus extracts from the US Federal Register and Congressional Record. Participants were required to use their retrieval systems to rank the documents on their relevance to each topic, and to send the IDs of the top 1000 documents to NIST for assessment.

An example of a TREC topic statement: Topic 261

```
<top>
<num> Number: 261
<title> Topic: Threat posed by Fissionable Material

<desc> Description:
Does the availability of fissionable material in the
former states of the Soviet Union and its susceptibility
to theft, pose a real and growing threat that terrorist
groups/terrorist states will acquire such
material and be able to construct nuclear weapons?

<narr> Narrative:
Under the terms of the strategic disarmament treaty with
the U.S., the states of the former Soviet Union have been
dismantling 2000 warheads each year.  From each warhead a
shiny sphere of plutonium is extracted.  These spheres,
called "pits", are the elemental cores of a bomb.  In addition,
other forms of plutonium are scattered over the former Soviet
Union in institutes, laboratories, plants, shipyards and
power stations.  Disgruntled employees, who are often underpaid
or paid irregularly have access to the plutonium.  This worries
leaders in other countries.  Enriched uranium, an alternate fuel,
is harder to come by because it is stored in well-guarded military
facilities, but it is easier to turn into a bomb.  The Russians
have denied that it came through or from their country, but German
authorities believe that it did.  Any item which speaks to failures
in the safeguarding of nuclear material or to black-market operations
in nuclear material, or to efforts of terrorist groups or terrorist
states to acquire such material would be relevant.
</top>
```

The union of the top N[7] documents from each run for a topic formed the judging pool for that topic. Documents in the topic pools were judged for relevance by a team of retired intelligence

[7] Typically $N = 100$.

assessors who marked each (topic, document) pair as Relevant or Irrelevant.

Using a file of these relevance judgments (known as `qrels` for some reason), the performance of each run could be measured. Runs were evaluated down to rank 1000, assuming unjudged documents to be Irrelevant.

At any point in a ranking you can calculate:

- *Precision* – the proportion of retrieved documents which are Relevant, and
- *Recall* – the proportion of the Relevant documents which have been retrieved at that point.

You can then plot precision versus recall at each point where a relevant document is retrieved. NIST did this and averaged across the topics to give a precision-recall graph for each run. They also calculated other measures such as P@10 (precision at ten documents retrieved), and MAP (mean average precision – the area under the precision-recall curve.) An example precision-recall plot appears on Page 22.

The resulting aggregation of documents, topics, and qrels forms a *test collection* which can be used for testing and training retrieval systems after the conference. Over the years, the amount of TREC training data accumulated. Participants were often reminded that, due to over-fitting, it's easier to achieve great retrieval results on past data than in the blind conditions of a new TREC ad hoc.

There were two main participation categories in TREC ad hoc: Manual – where a human turned the topic statements into queries for the retrieval system, and Automatic – in which the conversion was automatic, or where the retrieval system accepted the topic statement as the query. Subcategories of Automatic participation related to how much of the topic statement was used. Some used the full topic text, others only title plus description, others title only.

You were only permitted to attend the TREC conference, held on the NIST campus in Gaithersburg, MD, if you submitted at least one run for evaluation, or if you were a government sponsor. I recall Donna naming a company which had submitted a run but then successfully pleaded to have its results withdrawn. I also remember one company being exposed as having "cheated" – i.e. they manually intervened in an allegedly Automatic run.

On arrival at the conference, participants were presented with a huge binder containing all the results, and papers submitted by each of the participants to describe their approach. University participants were expected to describe their methods in full detail, but commercial companies were allowed to suppress proprietary details.

I couldn't wait to participate.

1998: Members of the ACSys WAR team: Jason Haines, Paul Thistlewaite, David Hawking, Nick Craswell.
Photo: ACSys.

2.4 ∗The ACSys Cooperative Research Centre

In late 1993, ANU and CSIRO successfully bid for a co-operative research centre (the Advanced Computational Systems CRC – ACSys), which involved Fujitsu, Sun and DEC as commercial participants. Paul Thistlewaite was the leader of a foundation project known as PASTIME (Parliamentary Sound, Text and Image Environment) which partnered with the Parliamentary Information Systems Office (PISO). It used the newly developed CERN `httpd` web server and NCSA `Mosaic` browser to dramatically improve access to parliamentary information such as Hansard transcripts.

Over time, Steve Ball and Jason Haines joined Paul on the project, Nick Craswell became a PhD student, and I was seconded 40% of full-time to work on text retrieval technology.

I started work on converting PADDY into PADRE (Parallel Document Retrieval Engine), which required quite substantial rewriting and the introduction of new operators in order to better handle the challenges posed by TREC topics. Paul Thistlewaite worked on code to convert the TREC topics into PADRE queries and I developed a set of manual queries. Our TREC-3 paper in 1993[8] valiantly attempts to sell the advantages of index-less search on a parallel machine. Alan Smeaton from Dublin City University told me that PADRE's capabilities seemed like a solution in search of a problem. I fear that he was largely correct.

I performed the runs on the 512-node AP1000 at Fujitsu Laboratories in Kawasaki, Japan. I could book two-hour slots during which I had exclusive use of the machine. The slots were rigidly enforced – your job was terminated the instant the slot ended, even if only another few seconds was needed to finalise the results. Only two separate runs were possible within a slot since it took 45 minutes to load the data onto the nodes of the AP1000. The process was complex and somewhat stressful. I remember my wife Kathy Griffiths sitting with me and assisting.[9]

The PASTIME project was succeeded by TAR (Text And Related) and WAR (Web And Related projects). This line of research continued until the end of ACSys in September 2000. A focus was on achieving linear scaling – Given a retrieval problem N times as large as the original, can you achieve the same response time if you deploy N times as many compute nodes as for the original? A related question is whether you can reduce response time by a factor of N if you multiply the number of nodes by N. That depends upon how much of the computation is inherently sequential, i.e. not amenable to parallelisation. Amdahl's law[10] gives the speed-up you can achieve from parallelism for a computation whose inherently sequential part represents a proportion s of the overall computation. Even with infinite parallelism the maximum possible speed-up is $\frac{1}{s}$. For example, if half of the total computation is inherently sequential ($s = 0.5$) the maximum possible speed-up through parallelism is only 2.0.

Assuming that the corpus of documents is partitioned by document, i.e. that documents are distributed across nodes, text retrieval is what is known as an "embarrassingly parallel" problem. Personally, I felt no embarrassment – this type of problem is ideal for a parallel machine.

There are several flies in the embarrassingly parallel ointment. If there is an imbalance across the nodes, the speed of response is determined by the slowest of the nodes. Load imbalance can occur if the quantity of text differs from one node to the next or if query words occur in the text held by one node far more than others. In general it is not possible to achieve perfect balance. To guarantee response time it may be necessary to impose time-outs, cutting short computation on slow nodes.

The second problem is that of communication. The time taken to send partial rankings back to the front-end is not scalable. Indeed, the more nodes, the greater the amount of data which must be communicated. Even worse, if nodes are perfectly balanced, all nodes will transmit simultaneously and serious network contention is likely. Fortunately, the AP1000 had an unusually high communications bandwidth to compute power ratio.

Finally, the front-end must aggregate and sort all the partial result lists into a single list. This process is essentially sequential, bringing Amdahl's Law into play. Furthermore, the scale of the aggregation problem increases with the number of nodes.

[8] `http://david-hawking.net/pubs/HawkingT94.pdf`

[9] Kathy remembers this happening at around midnight.

[10] `https://en.wikipedia.org/wiki/Amdahl%27s_law`

These three problems have been faced by modern web search companies such as Google, Bing and Baidu, where a single query may be processed on a network of thousands of servers.

The focus of PADRE soon changed away from index-less search. By itself, it wasn't enough to show quasi-linear scale-up with number of nodes – we wanted to show that the performance of PADRE on a single node was competitive with the retrieval systems of other TREC participants, and then show linear scale-up. In 1997 I presented progress on this front at the European Conference on Digital Libraries (ECDL 97), held in Pisa Italy.

ECDL 97, Pisa, Italy

I attended the European Conference on Digital Libraries in Pisa to present more credible claims for the scalability for PADRE running on a COW (cluster of workstations). It was the best catered conference I have ever attended. At breaks, white-tied, besuited baristas made excellent espresso behind a marble bar.

The conference dinner was held in a villa some distance outside Pisa. I'm sure I counted 14 courses, all of them exquisite, and most of them accompanied by an alcoholic beverage. The "evening" ended downstairs with espresso and grappa, before buses returned us to our hotels at about 2am. After stirring renditions of *The Fields of Athan Rye* by two Irish colleagues, Francis Crimmins (who I later recruited to CSIRO) and Páraic Sheridan, I resisted their exhortations to join them at an Irish pub which happened to be open at that hour.

The next day, the Irish contingent was sorely depleted. One of them arrived at lunchtime to show the flag. He was pasty faced, and his eyelids needed to be propped open with matchsticks to show off his bloodshot eyes. In a cruel rebuff of his good intentions, there were no sessions after lunch that day! The conference was over.

After the conference I made a brief visit to Florence to check out a couple of minor art works ☺ and to purchase a leather bag for Kathy. The stall-holder told me in shocked tones of the death of Princess Diana.

A day or two later I was scheduled to lay over in Heathrow Airport for a few hours prior to my flight back to Australia. Kathy encouraged me to head into Kensington Palace to see the floral tributes and the BA flight crew assured me that the tube was always quiet on a Sunday evening, so I would easily get in and out in time for my flight home.

They were right as far as the Picadilly Line was concerned but transferring to the District Line for High Street Kensington made me question the wisdom of my excursion. The train was grossly overloaded and seemed to keep blowing a circuit breaker. Each time, the lights went off and the train stopped. Fortunately, we eventually made it, but special arrangements were in place. All the ticket barriers had been removed to avoid the crush, and police were enforcing one way pedestrian flow outside the station.

The volume of flowers, cards, and mementoes outside the palace was absolutely astonishing, and the feeling of grief among the people was palpable. I bought a tabloid newspaper on the way back and found that, out of 132 pages, something like 128 were about Diana.

Over the life of ACSys the scale of desktop hardware underwent dramatic improvement, to the point that quite large text corpora could be indexed and searched without the need for parallelism.

2.5 *1994: First TREC Participation

I attended TREC-3 in November 1994. It was my first visit to the USA and it was quite an eye-opener. First of all, the pilot flying us from Los Angeles to Washington Dulles frightened me by anouncing that we would be, "taking off momentarily."[11] A cabin steward was totally unable to understand my request for a "tomato juice." "Say what???". Then I read a newspaper report of a weekend shooting party at a farm in Idaho, in which more than 70,000 rounds of ammunition were fired from guns ranging from rifles to machine guns to heavy military weapons. The police were called by frightened neighbours but the only offence being committed was making excessive noise on a Sunday.

As was my habit, I tried to travel from the airport to my hotel by public transport. I waited

[11] Obviously he meant "soon". In my linguistic community, it always meant, "for a brief instant."

an hour for the Washington Flyer (bus) to take me to the West Falls Church metro station, waited for a considerable time for an Orange line train to Metro Center then changed to a Red line train to Shady Grove. The metro service wasn't anywhere near as frequent as those in London or Paris but the cavernous stations with their subdued lighting were a definite improvement. Hours later, at Shady Grove, I thought about trying to walk to the Gaithersburg Holiday Inn, but was put off by the shortage of sidewalks and a lack of knowledge about how to get there. The taxi ride took ages, and cost a lot, surprisingly cash-only. The driver was unambiguous in communicating that a tip was essential and that more than my initial offer was required.

At check-in, the clerk told me that my accent was very cute, and gave me a card telling me that the policy of the hotel was "aggressive friendliness." **Have a nice day!!! ... or else!** My room was absolutely enormous, and had a toilet whose handle had to be pumped up and down to achieve a swirling flush. The bath was gargantuan but the faucet delivered a niagara of water, sufficient to fill it in a jiffy. There was no coffee worthy of the name within 30km of the hotel!

Donna Harman had told me that Australians from Melbourne were already TREC participants and electronically introduced me to Ross Wilkinson, from RMIT. He and I met for breakfast before the first day of the conference. Ross saw some people he knew and we sat at their table. He introduced me to Stephen Robertson and Micheline Beaulieu (City University, London), Karen Spärck Jones (Cambridge University) and Alan Smeaton (Dublin City University). I had no idea that Stephen and Karen were exalted figures in the information retrieval world and proceeded to explain to them how I thought that retrieval should be done. How very polite they were! How naive I was!

Left: Stephen Robertson and I "bunked off" from the Information Retrieval Festival at the University of Glasgow in 2007. We climbed Goat Fell on the Isle of Arran. Right: Karen Spärck Jones in 2002. Photo: By University of Cambridge, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=4734533

On arrival at the NIST campus, several features stood out: The US flag fluttered on a tall flagpole, and the very large campus was full of ground hogs, deer, and goose excrement. Hundreds of participants gathered in a large auditorium in the main administration building, waiting breathlessly to learn how their results stacked up against others.

TREC-3 was a great conference at which to start one's information retrieval research career. The SMART retrieval system from Cornell had "won"[12] each of TREC-1 and TREC-2 but, in TREC-3, SMART was soundly beaten by InQuery from UMass at Amherst, and by Okapi/BM25 from City University, London. Amit Singhal[13] and Cornell colleagues subsequently showed that City's per-

[12]The organisers kept emphasising that TREC is not a competition.

[13]Later Head of Search Quality at Google

formance gain was due to better document length normalisation and that by incorporating it into SMART, Cornell was again highly competitive.[14]

Ross Wilkinson (right) and I attempt to strangle fellow researcher Andrew Turpin from the University of Melbourne. SIGIR 2006 conference, Seattle WA. *Photo: Alistair Moffat*

Our ACSys runs performed poorly but I soaked up all three days of TREC like a sponge. There wasn't a presentation I didn't get ideas from. I had the privilege of listening to a crusty extended monologue from Gerard Salton of Cornell University during what was supposed to be a discussion of a possible passage retrieval track in the next TREC conference. I met Bruce Croft, James Allan, and Jamie Callan from UMass, Jacques Savoy from the University of Neuchâtel, and chatted with many researchers about all sorts of different approaches to retrieval: probabilistic, vector space, Markov models, inference networks, n-gram indexing, Rocchio feedback, and latent-semantic models.

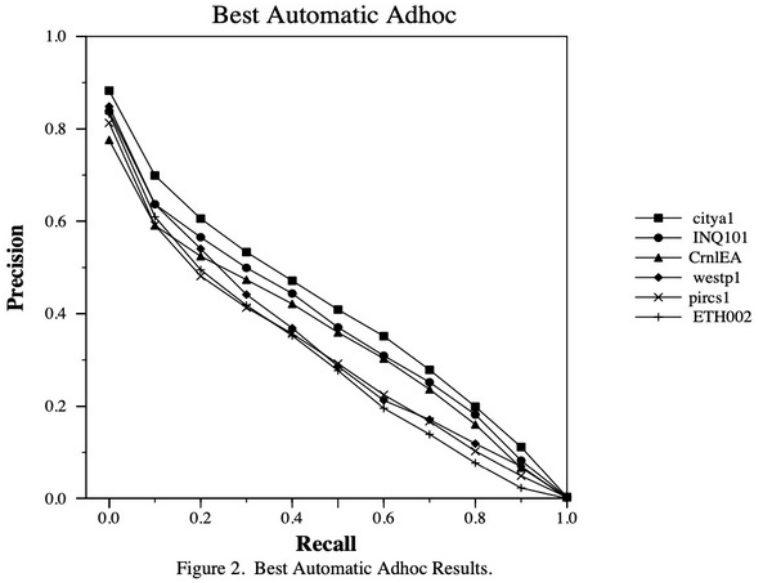


Figure 2. Best Automatic Adhoc Results.

The TREC-3 leader board for the Automatic ad hoc participation category. City University, UMass, Cornell fill the first three places. ANU, along with dozens of other participants, is nowhere to be seen. *Taken from the TREC-3 Overview.*

[14] `(https://ecommons.cornell.edu/bitstream/handle/1813/7186/95-1529.pdf`. Length normalisation – how to correct for the fact that a particular number of occurrences of a query word is more significant in a short document than a long one.

1994: Clockwise from top: NIST administration building with flagpole; Reminder of Halloween in Washington DC; Corridor in NIST building where TREC workshops were held.

1994: Top to bottom: Washington Dulles airport (still under construction) with scissor-lift "mobile lounges" to transfer people from the island terminal to the main building; A pleasant lunchtime setting outside the NIST cafeteria; A groundhog burrow on the NIST campus.

A reception held at the Gaithersburg Holiday Inn paid no mind whatever to dietary constraints. Vegetarians fought over a single bowl of carrot sticks, the only concession to their diet, and mostly drove off elsewhere to find something to eat. Subsequent TREC banquets in the Cracked Claw in Gaithersburg did at least provide vegetables, salad and french fries in addition to huge mounds of crabs.

At the TREC-3 poster session, I was surprised to find a poster on n-gram retrieval by an environmental research institute. An American informed me, "that's just a front for the CIA!"

At the wrap-up session at the end of the conference, suggestions were invited on what should happen in future TRECs. I tentatively asked whether it was time to consider increasing the scale of the document set, since electronic document collections around the world were growing rapidly. Dave Lewis (who later became a friend) jumped up and shouted, "That's the guy with eight gigabytes of RAM!"[15] I wasn't put off and over the years we succeeded in increasing the size of TREC test collections by two orders of magnitude.

2.6 *1995: TREC-4 – Proximities and Z-mode

Probabilistic retrieval systems, like InQuery (UMass), Okapi/BM25 (City University, London), and our TREC-3 system, attempt to rank documents in order of decreasing probability of relevance. They generally don't estimate probabilities but instead use scores which produce the same ranking as probablities would. The relevance score of a document is estimated by taking into account the presence or absence of query terms in the document, their relative importance, and their frequency of occurrence. tf represents the frequency of occurrence of a term in a document. df is the document frequency – the number of documents in the collection containing the term. A term's importance is reflected by idf – the inverse document frequency. The fewer documents a query term occurs in, the stronger the relevance signal it confers. A query term occurring frequently in a candidate document but infrequently in other documents (i.e. high $tf.idf$) is a strong indication that the candidate is relevant.

In such systems all the query words are generally treated identically, even if they are synonyms or overlap in meaning. Furthermore, because a document's total score is the sum of the component scores due to each term, the highest ranked documents don't necessarily contain all of the query terms. People using early web search engines such as Alta Vista to search for `david hawking` (was it really only me who searched for that?) found that top ranked results related to Stephen Hawking and contained no occurrences of the word `david`.

In TREC-4, we tried to overcome these deficiencies. We modeled TREC topics as a set of concepts and hypothesised that a relevant document should have evidence for the presence of all of the concepts. For example, the topic "What is the economic impact of recycling tyres?" contained three concepts: "economic impact", "recycling", and "tyres". Evidence for the presence of the economic impact concept might be "dollars", "financial return", etc.

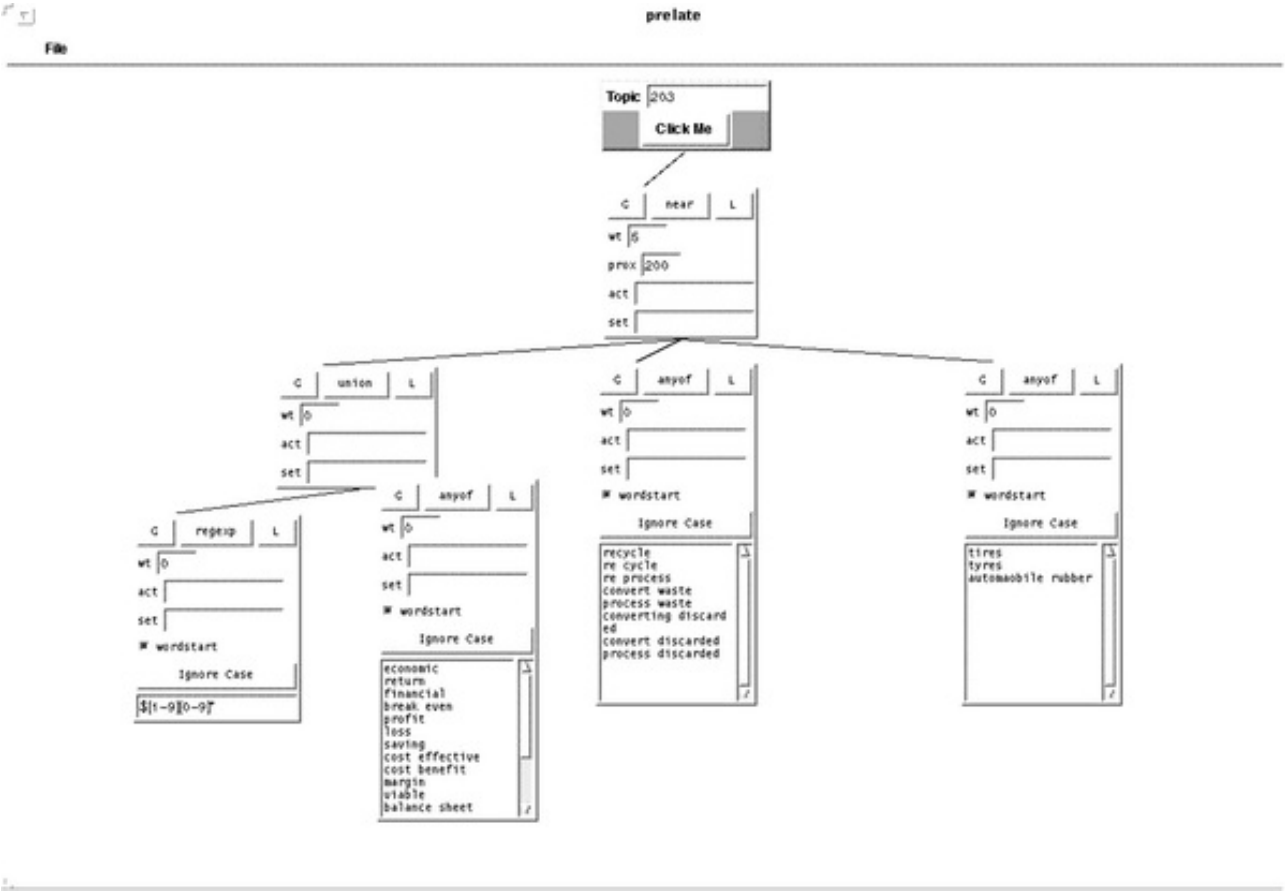
Initially our queries were in conjunctive normal form – A number of concepts represented by ORs linked at the top by ANDs.

```
(dollars OR "financial return") AND (recycling OR reprocessing) AND
(tyres OR rubber).
```

We then hypothesised that evidence for each of the concepts should really occur within close proximity, i.e. replace the AND operators with NEAR operators. (Many of the documents covered several different topics. A reference to `rubber` occurring 40 pages after `dollars` and `recycling` was considered unlikely to signal relevance.)

We created a graphical interface called `PRELATE` (PADRE Retrieval Language Topic Editor) to assist with creation and display of queries in this form.

[15] I.e. about 512 times as much RAM as he had on his machine.



The TCL/Tk *PRELATE* interface to help with building manual queries for PADRE. ***Taken from Hawking and Thistlewaite, Proximity Operators – So Near and Yet So Far https://david-hawking.net/pubs/HawkingT95.pdf.***

Boolean queries misunderstood by many search engine users.

From the beginning, the Alta Vista web search engine supported Boolean operators, as well as proximities, +, and −. However, an analysis of a billion queries by Craig Silverstein et al in 1998[a] showed that only 20% of queries contained any operators. At SIGIR 1998 Monika Henzinger (one of the authors) reported that many queries using operators contained obvious errors, and that Alta Vista had found that many users misunderstood operators like AND. They thought that `cat AND dog` meant, "Give me documents containing `cat` AND documents containing `dog`." It actually means, "Give me documents containing both `cat` and `dog`."

[a] `https://www.hpl.hp.com/techreports/Compaq-DEC/SRC-TN-1998-014.pdf`

For our Manual TREC-4 submission we composed extremely long queries after consulting with subject experts. In the topic, "What is the role of vitamins in human health", the two concepts "vitamins" and "human health" were each represented by hundreds of words, e.g. vitamins by `vitamin`, `riboflavin`, `niacin`, ...

To speed up full text scanning I devised a version of the BMG[16] algorithm which allowed for multiple alternate patterns. You could search for `vitamin|niacin|riboflavin` in a single pass through the text. I presented this at a Friday lunchtime meeting of the ANU-Fujitsu CAP project. I was proud of the achievement but conceded that the speed-up due to BMG reduced as the number of alternates increased. Andrew Tridgell said, "You could use bigrams!" (I.e. make a skip table for character pairs rather than for single characters.)

Over the weekend, Andrew coded up `bmg2`, implementing trigram as well as bigram skip tables, and including several tricks to reduce the amount of memory required. It turned out that bigrams produced a dramatic speed-up and that trigrams weren't really necessary. With Andrew's permis-

[16] See Page 15.

sion, we used `bmg2` in our TREC-4 submission and for smaller purposes for decades thereafter. It's open source of course!

We made some runs on the TREC-4 task using a fixed proximity. I.e. evidence for each of the concepts had to occur within say 50 words of each other. Then, one night at around 3 in the morning, while drinking coffee and eating Tim Tams in the CS&IT tea room, Paul Thistlewaite and I devised a scheme called "Z-mode", where the proximity limit was relaxed and each "concept span" was scored based on the number of intervening words. "Large profits from recycled rubber" would score highly because the concept span has only one intervening word, "from".

Paul and I submitted TREC-4 runs based on Z-mode. Interestingly, Charlie Clarke and Gord Cormack from the University of Waterloo, Canada also (independently) used a proximity based approach. Neither group appears in the leader boards for either Manual or Automatic approaches.

However, we also participated in the "Database Merging" task with our Z-mode method. In this task, the TREC ad hoc corpus was divided into ten sub-corpora. Queries were to be run over each sub-corpus independently and the ten result lists merged into a single list. The merged list was then compared to the single list obtained by running the same method over the corpus as a whole. The other participating systems found this a significant challenge because idf values can vary dramatically from one sub-collection to another. For example, "algorithm" may be highly discriminating in a newswire sub-corpus but hardly significant in a sub-corpus of computer articles. As a result, scores from the different sub-corpora are not compatible. Z-mode made no use of idf and its scores are compatible across sub-corpora. The graph on Page 27 shows that Z-mode runs achieved best performance and were identical between the merging run and the baseline.

Owen de Kretser at the University of Melbourne pursued similar distance-based ideas in a PhD thesis supervised by Alistair Moffat, which I examined in 2000.

We continued our argument for the use of larger document corpora and were successful to the extent of persuading NIST to run a trial of a Very Large Collection (VLC) track in TREC-5, using twice as much data (4GB) as before.

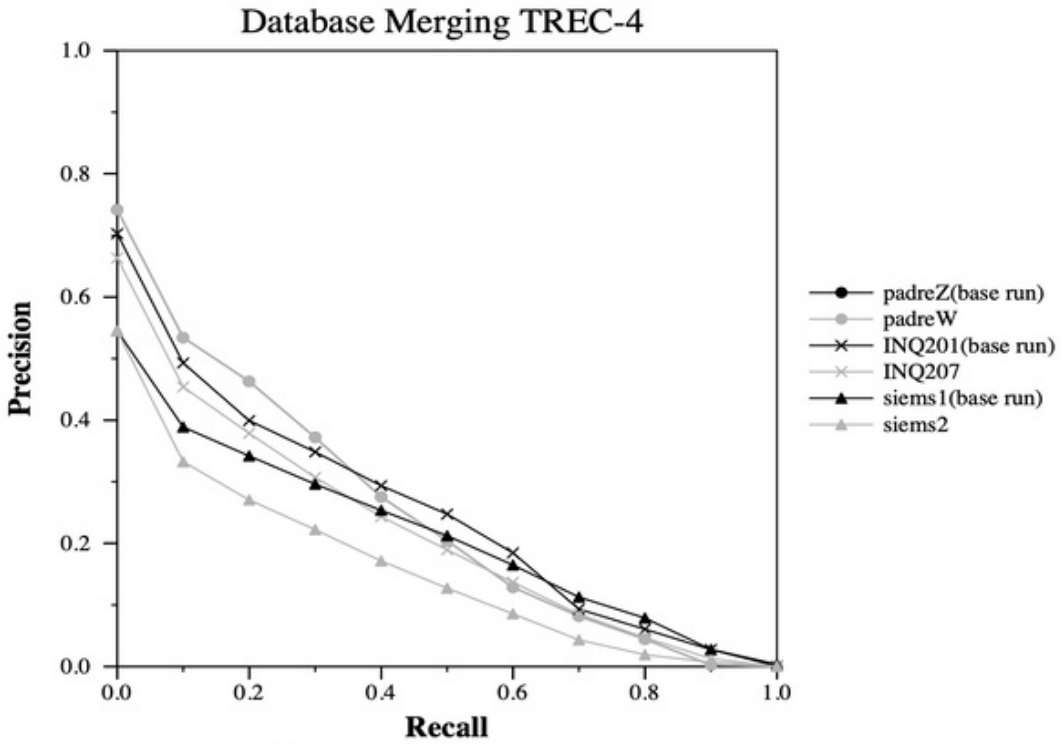


Figure 11. Results of TREC-4 Database Merging Track.

The TREC-4 leader board for the Database Merging participation category. The plot for padreZ is invisible, being overlaid exactly by the identical plot for padreW. *Taken from the TREC-4 Overview.*

2.7 *1996: TREC-5

Our report on TREC-5 participation[17] shows us continuing to work on scoring methods based on lexical distance. Paul and I wrote a technical report on this type of scoring,[18] generalised to support partial spans and different scoring formulae. I implemented this in PADRE. I also tried to determine whether distance-based scoring was capable of performing better than traditional $tf.idf$ methods. I.e. whether our performance in TREC-4 had been held back by inferior queries. I obtained and translated the University of Waterloo queries and ran them through PADRE, confirming that our results were at least as good as theirs had been. I then tweaked queries until I achieved results better than the TREC-4 "winner."

I played around with expanding queries based on term implication (i.e. the presence of one term often implies the presence of another), and also implemented mechanisms for matching numerical quantities and for representing a US context. At the same time, we experimented with server selection methods in the Database Merging track, and participated in a task requiring retrieval of relevant documents from an OCR-scanned material. Naturally we participated in the VLC pre-track.

Assumed US context

In working with people and organisations in the US, I often became aware of an implied US context. Once, when coordinating time-critical activities with a US researcher, she said, "Oh that won't work because you'll be celebrating July Fourth." I said, "No, I'm in Australia. We don't celebrate July Fourth." To which she replied, "When do you celebrate July Fourth?"

In the meantime, Paul worked on automatic query generation and achieved good gains – we scraped into an expanded leaderboard. Most of our work made use of Peter Bailey's "super-dictionary" indexes kept on the disks of ANU's AP1000.

The years of TREC-5 and TREC-6 were a period of furious research activity. I was now 100% seconded to the ACSys CRC and working full-time in a fever of excitement. Despite this our TREC results were a little disappointing, perhaps because we spread ourselves too thinly. In TREC-5 we started using statistical tests (t-test) when comparing runs. This eventually led to the realisation that there was often no real performance difference between the first half-dozen leaders in the TREC tasks.

2.8 *Creating the TREC Very Large Collection

Once the TREC-5 conference was over I started working flat out on accumulating documents to make up a VLC corpus ten times as large as that used in the main ad hoc task. I relied heavily on Paul Thistlewaite's contacts in parliament and government, and on friendships I had made at TREC. (Paul accused me of being "promiscuous" since I would talk to anyone. ☺)

I became Coordinator of the forthcoming VLC track and was invited to join the TREC Program Committee (PC). Meetings of the TREC PC were traditionally held at the house of Ellen Voorhees (co-co-ordinator of TREC), her husband Chris Buckley, and their young son Nathan. Donna Harman contributed casseroles to a meal, and the highlight was the selection of weird icecreams chosen by Chris and Nathan. Think bacon and peanut butter!

I later visited Mark Sanderson at the University of Glasgow and gave a seminar. Mark introduced me to Kate Marsden at the Financial Times (London) and to his University of Glasgow colleague Jon Ritchie. Jon was able to persuade Caledonian Newspapers Ltd to provide several years of articles from the Glasgow Herald. I visited Kate in a Financial Times building in central London and afterward she made the case to management for releasing three years of Financial Times articles. A couple of weeks later she emailed me and Mark, "Thunderbirds are go!"

While in Glasgow, Mark and his wife Pam took me on a weekend walk in Glen Coe. We walked up a hill across the valley from the Buachaille Etive Mor. Armies of people were streaming up the

[17] https://david-hawking.net/pubs/HawkingTB96.pdf
[18] https://david-hawking.net/pubs/HawkingT96.pdf

Buachaille, determinedly "bagging" a secondary summit which had been recently added to the list of Munros – the Scottish peaks exceeding 3000 feet in height.

1997: Me and Mark Sanderson at Glen Coe, Scotland. *Photo: Pam Sanderson*

After descending our slightly-short-of-a-Munro hill, we lunched at the isolated Kingshouse Hotel, where we met a former student of Mark's. He had travelled all the way from Plymouth for the long weekend, to camp by the creek near the Kingshouse. He begged Mark for a ride back to Glasgow. Mark agreed but we had to walk a few kilometres to fetch the car. As we drove into the Kingshouse carpark we saw the strange sight of a man (the former student) draped head to foot in wet weather gear, face covered with mesh, jumping up and down, slapping himself. He wrenched the door open and dived into the car before it had even stopped. As he disrobed we could see that he was covered with red blotches the size of 20 cent pieces. Bloody midges!

The report of the TREC-6 VLC Track[19] lists more than 20 organisations who donated data to the VLC. Data included Hansard reports from Canada and Australia, Australian legislation and legal judgments, and a wide range of web sites. In addition, we used Usenet News data archived by Gord Cormack and Rob Good at Waterloo, books from Project Gutenberg, and all of the TREC CD-ROMs.

Jason Haines, Tim Potter and Nick Craswell worked on converting all the data into TREC format.

2.9 ∗1997: TREC-6 – BM25 and Relevance Feedback

Although the distance-based scoring methods were very appealing and had the valuable property of being collection-independent, we had not found a good way of automatically deriving appropriate queries from the TREC topics. Accordingly, our TREC-6 runs reverted toward the methods others were using. We also moved away from the AP1000, instead using a single workstation or a cluster of workstations (each with local disks.) I completely re-engineered PADRE (PADRE97) to suit the new environment and to implement a variant of BM25.

I found that using BM25 more or less doubled the MAP scores achieved by the arbitrary $tf.idf$ function I'd used in TREC-3. It seemed that having a solid theoretical foundation was a winner!

[19] https://david-hawking.net/pubs/HawkingT97.pdf

Even with this big gain, we would have still been a long way behind the pace, because other systems used *pseudo relevance feedback*. A relevance feedback mechanism calls on a human to judge the relevance of top ranked documents. It then mines the relevant documents for additional search terms which are added to the original query or used in place of it. Weights are usually associated with the terms. The expanded query is then run to produce a new and usually improved ranking.

In *pseudo relevance feedback*, there is no human in the loop – the top k[20] documents are assumed to be relevant. Even irrelevant documents may supply search terms to assist in retrieving relevant ones.

To illustrate the method, imagine that the original query were `north sea oil exploration`. A relevance feedback mechanism might discover that top ranked documents contain higher than usual occurrence rates of words like: `offshore`, `platform`, `drilling`, `statoil`, `BP`, `petroleum`, and so on. Using a substantially expanded query, the final run may retrieve a lot more relevant documents, and even improve precision early in the ranking.

PADRE97 implemented a passage-based pseudo relevance feedback method. Instead of extracting feedback terms from whole documents, they were extracted from "hot spot" passages surrounding query term occurrences in the text. Selection and weighting of feedback terms used a formula due to Stephen Robertson. In training, despite a few topics being seriously harmed,[21] the new pseudo relevance feedback method showed dramatic gains on the Automatic Short Topic tasks from TREC-3, 4, and 5.

For TREC-6 Nick Craswell developed a TCL/Tk tool `Quokka` (more sophisticated than `PRELATE`) to help with developing manual queries. My memory is a little hazy but I think that `RAT` (Relevance Assessment Tool) was derived from `Quokka` and that Jason Haines was involved in creating it.

In the blind conditions of the TREC-6 campaign, we didn't blitz the field in the Automatic Short category but we did manage to top the leader board in Automatic Long. We achieved best MAP, best recall and best precision@20 in that category. We also made it to the leaderboard in the Manual ad hoc category. Almost certainly no statistically significant difference between us and the others but we were delighted.

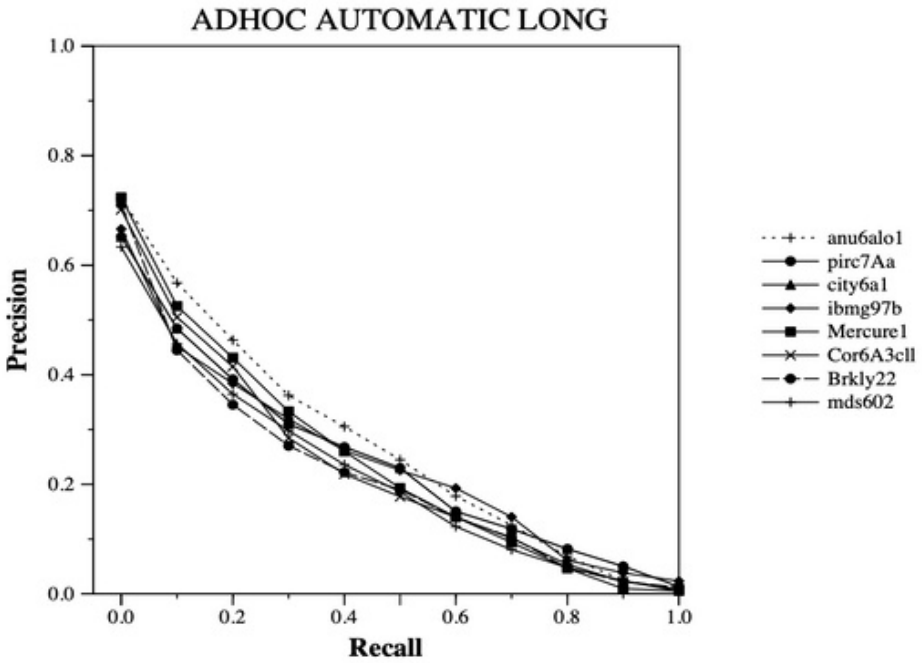


Figure 5: Recall/Precision graph for the top eight automatic ad hoc runs using the full topic.

The TREC-6 leader board for the Automatic Long participation category. *Taken from the TREC-6 Overview.*

[20] Typically k is ten or twenty.

[21] A normal occurrence with relevance feedback. In Bodo von Billerbeck's PhD work, which I examined, he showed an example of a topic where the addition of one specific feedback word to the query dramatically improved performance, but adding any of the other feedback terms degraded it.

Nick Craswell worked on ACSys submissions in the Filtering Task. In filtering, the documents in the TREC corpus are treated as a stream of incoming messages, and the retrieval system assesses relevance message by message, using a score threshold to decide whether to accept a document for reading by a recipient or reject it. Nick used a Generalized Reduced Gradient nonlinear optimisation method provided by Microsoft Excel 97 to tune PADRE parameters and set thresholds based on training data from earlier TRECs. This method worked quite well.

2.10 ∗1997: TREC-6 VLC track

To persuade NIST to agree to running the VLC track, ACSys committed to distribute the data and also to provide resources for relevance judging. We copied the formatted (and compressed) VLC data on to data tapes (DDS-2 format I think) and mailed them to fourteen groups. The tapes also included a 10% baseline sample of the VLC.

Each group was required to sign a data permission agreement, committing them to use the documents only for research into information retrieval or natural language processing methods, and to delete all copies of particular documents within the corpus, if requested to by ACSys. This was our protection against included documents being in breach of copyright or containing defamatory material.

John O'Callaghan was CEO of ACSys at the time and, since he would sign for ACSys, he sought advice from a law firm about my draft of the agreement. The lawyer came back with a few small changes.

> **Dave:** But you didn't say whether you were happy with the agreement overall! The real question is, "Will John go to jail?"
> **Lawyer:** It's unlikely that John will go to jail.

With my gratitude and admiration, John signed the agreements on that basis. Fortunately, no request or demand was ever received for removal of documents either in the VLC or in any of the subsequent collections distributed by ACSys or CSIRO.

Topics were the same as those used in the TREC-6 ad hoc tasks but, naturally, new judgments were needed.

VLC was large enough to cause obvious violation of one of the fundamental assumptions in TREC ad hoc – It was no longer safe to assume that unjudged documents were Irrelevant. We decided to focus on early precision – what proportion of the first 20 documents retrieved were relevant. This seemed justifiable given that only a very tiny percentage of web searchers scan beyond the first 20 results. Experienced TREC people argue that evaluation down to depth 1000 provides valuable extra information about a system's ability to discriminate Relevant from Irrelevant documents. In my opinion, a better way to gather more information is to invest judging resources into shallow-judging a much larger set of topics.

We judged every document in all the submitted top-20s, so unjudged documents were not an issue. This had the downside that the VLC collection was not re-usable in the TREC sense – future runs are likely to return unjudged documents which have a non-negligible probability of being relevant. However, we also demonstrated that the cost of making relevance judgments for top-20 lists was quite affordable, meaning that follow-up experiments could be meaningfully conducted with a small investment in judging.

ACSys funded the casual employment of relevance assessors, Deborah Johnson, Sonya Welykyj and Josh Gordon. They used the RAT to perform assessments. The RAT displayed the topic to be assessed and asked the assessors to specify the concepts which they thought essential to the topic. Each concept was associated with a colour and a list of words, phrases or part-words which might provide evidence for that concept in a document being assessed. Assessors could enter the evidence terms by typing or by highlighting fragments in the text of a document being assessed.

When a document was displayed, every occurrence of an evidence fragment was highlighted in the colour of the associated concept. As assessment progressed, the stock of possible evidence built

up, and made judging very easy, mostly just skipping to fragments of the text where all the concept colours were present. It made sense to show the assessors the documents in order of increasing length. It was easy to spot previously unmarked evidence in short documents and by the time long ones were encountered there were enough evidence markers to rely on.

As an experiment, Paul got one assessor to work from longest to shortest. As luck would have it, the very first document encountered was longer than several concatenated PhD theses and covered a multitude of topics. Not having built up a reliable set of evidence markers, it took the assessor eight hours to judge that single document.

Furious Coding

Avoiding bugs in a large scale text retrieval system is of major importance. Because you don't know what the correct answers are, and because it is impractical to scan thousands of retrieved documents looking for query words, it is possible that small errors in compressed index structures or in the compression code might go undetected and silently harm results. *It was a recurring fantasy of mine that I would discover a terrible bug, fix it, and see a massive jump in MAP!*

At one stage, close to TREC submission deadline, I suspected that there were bugs in my indexing code. Accordingly I sat down and wrote an elaborate program to validate the index and to meaningfully display its contents. I wrote a thousand lines of C in one day and got it working. I can't remember for certain whether it found any errors in the index but I assume it did, and that they were fixed but didn't dramatically affect the results.

It comprised a single source file `index_check.c`. After I had written it and used it, I asked myself where in the source directory structure it should be filed. Having found the right place I noticed that there was already a file there called `check_index.c`. ☺

Yes, I'd forgottten that I'd already written a program to do the job!

Seven groups participated, including ANU/ACSys. The groups used a variety of hardware configurations ranging from single workstations to clusters of up to eight. We collected a lot of data from each participant, including the cost of hardware. We compared VLC versus baseline for precision@20, average query processing time, data structure building time, data structure size, and even "gigabyte-queries per hour per kilodollar". The latter was an attempt to even up the playing field for participants using cheap hardware. The University of Waterloo, which had efficient code and used very cheap PCs, scored almost two orders of magnitude better on this measure than the next best participant.

An interesting observation, which Stephen Robertson and I subsequently explained, is that precision@20 was substantially higher for VLC than for its 10% sample. Stephen and I worked on the very long *On collection size and retrieval effectiveness* paper[22] on and off for several years. Stephen came up with theoretical predictions based on signal detection theory and I ran experiments to confirm or refute them. We argued about models and I remember conducting an intense discussion through the window of a van taking him from the Gaithersburg Holiday Inn to Dulles Airport, continuing as the bus rolled out of the carpark! It was science at its best!

Stephen used signal detection theory to model retrieval. It was a cause of disappointment to me that you couldn't produce an ROC (Receiver Operating Characteristic) curve[23] for a retrieval system, but only for the retrieval system coupled with a query.

Query throughput for all of the systems was really slow compared to modern search engines. Several of the participating systems processed less than one query per minute. The focus of nearly all TREC participants was to achieve the highest quality ranked lists, regardless of the time taken.

2.10.1 *Showing Off

Many TREC participants seemed to find going beyond 2GB of data quite daunting. I attempted to show them that processing 2GB of text was very easy and could be done with quite tiny hardware.

[22] `https://david-hawking.net/pubs/hawking_robertson03.pdf`
[23] `https://en.wikipedia.org/wiki/Receiver_operating_characteristic`

As co-ordinator of the VLC track I was given a 30 minute speaking slot. I presented using a small Dell Latitude laptop running Linux. Before I commenced my presentation, I started a script to index the TREC ad hoc data, convert the topics into PADRE97 queries, run the queries, and evaluate the results. At the end of my presentation I showed the audience that the script had finished and presented the results. Some in the audience seemed quite stunned that respectable TREC results could be obtained with so little hardware in such a short time.

2.11 ∗1998: A Watershed Year

Things were looking bright in 1998. TREC-6 had shown that we had a retrieval system which could compete with the best TREC systems on both effectiveness and efficiency. Although we relied on techniques developed by others, we had made our own refinements and had pioneered the use of distance-based scoring. We had also developed effective tools for composing queries and for assessing relevance. We were the only group outside NIST who were doing relevance assessments for participants in shared retrieval tasks.

In April 1998, I gathered together eleven publications (not all of them refereed) and successfully supplicated[24] for a PhD by Published Work. I also changed employer.

When travelling internationally to attend a conference or meeting I always plan to arrive at least a full day early, to hedge against flight delays or jet lag. At TREC-6 Ross Wilkinson did the same and we spent the day exploring Washington DC's cultural institutions, such as the National Portrait Gallery. Ross had moved from RMIT University across Swanston Street to lead a team in CSIRO's Mathematical and Information Sciences Division (CMIS), under Rhys Francis. Ross's team was colocated with CITRI (Collaborative Information Technology Research Institute), a partnership between RMIT and the University of Melbourne headed by Ron Sacks-Davis and Kotagiri Ramamohanarao. Ross told me that there was an ongoing vacancy for an information retrieval research scientist in his group. I told him that I would have been interested except for three things: It required a PhD and I didn't have one yet; It was based in Melbourne and I didn't want to move, and; I was committed to the projects in ACSys.

Ross eventually told me that none of these problems were insuperable and encouraged me to apply. I did so, and soon found myself wheeling my belongings down the corridor to the CMIS wing of the CS&IT building, seconded 70% of my time to the ACSys CRC, and required to fly to Melbourne once a month.

2.11.1 International Conferences in Australia!

As noted elsewhere, the World Wide Web conference was held in Brisbane in 1998 and featured the very influential Google Anatomy paper.

For the first time in its history, the ACM SIGIR (Special Interest Group on Information Retrieval) conference was held in the southern hemisphere, in Melbourne Victoria, in 1998. This provided a boost for information retrieval research in Australia. Together with Justin Zobel, I ran a workshop on Efficiency in IR.

The Effective Retrieval Symposium: Canberra 1998

Paul Thistlewaite and I decided to capitalise on the Melbourne location of SIGIR and persuaded two "heavies" from the IR community to take a side trip to Canberra. Donna Harman and Stephen Robertson agreed to provide keynote talks at the very well attended *Effective Document Retrieval Symposium* which we organised and to which we invited many public servants and business people. Over the years, we were lucky enough to attract talks from many leading figures from the IR community: Ellen Voorhees (NIST), Bruce Croft (University of Massachusetts), Ricardo Baeza-Yates (then at the University of Chile), Mark Sanderson (then at University of Sheffield), Alistair Moffat (University of

[24] If you think "supplicating" indicates a shameful degree of grovelling, consider this: most PhD candidates "submit."

Melbourne), Justin Zobel (RMIT University and later University of Melbourne), Diane Kelly (then at University of North Carolina), and Peter Bruza (Queensland University of Technology). Many of these later talks were held under the banner of *IR and Friends*, a CSIRO talk and discussion series organised by Paul Thomas. Fueled by wine, cheese and orange juice provided by CSIRO,[25] these talks brought together people from CSIRO, ANU, NICTA and elsewhere around town.

2.11.2 Pressure to Commercialise

Up until this stage I was totally motivated by the research and engineering aspects of text retrieval. I wanted to rise to the challenges of retrieving useful documents as fast as possible. I loved hanging out with fellow researchers around the world and competing and/or collaborating with them to improve the standards of what was possible. I was so passionate about the research that I got into the habit of working 60–70 hours a week. I used to joke that WWW stood for "Weekends Without Work" and made sure that I had at least a couple of them every year. I found that it was almost impossible to do publishable research within the standard 38-hour CSIRO working week.

2008: Dave Abel, Deputy Chief of CSIRO Mathematical and Information Sciences Division, in typical pose – filling his pipe while working outdoors near ANU's CS&IT building.

I had no thought to try to turn PADRE into a money making opportunity, but both ACSys and CSIRO had strong commercialisation objectives. Indeed, CSIRO had a goal to earn about a third of its budget from "non-appropriation" sources. Since CSIRO overheads were of the order of 200%, this meant that each researcher needed to earn their own salary. Initially, there was no pressure on me because Ross Wilkinson's group had earned a large sum of money from the Public Records Office of Victoria (PROV) for advice on creation, storage, and management of electronic records – "Faxing the Future", as Rhys Francis described it. Problems arose the following year, when it was not possible to secure a similarly lucrative project. I immediately saw the potential benefit of developing useful software and licensing it to customers in return for an annual fee. Such an approach would address the need for external earnings and support longer-term, more interesting research.

It was expected that ACSys's commercial partners would pick up and commercialise the output of ACSys researchers. However, the Australian subsidiaries of Sun, DEC, Fujitsu and StorageTek had little ability to influence commercialisation decisions of their overseas parents. While ANU academics saw the necessity to work toward commercial outcomes, they seemed very reluctant to be actively involved in commercialisation.

[25] Thanks to Peter Bailey and later Carsten Friedrich!

Nick Craswell and I did manage to earn ACSys a small fee from DEC by conducting an evaluation of the effectiveness of DEC's PC version of the Alta Vista search engine. During a phone call with Alta Vista bosses in the US, we proposed applying the same methods to the Alta Vista web search engine. They said, "That's a great idea. We'll get some US PhD students to work on that."

During 1998, Paul Thistlewaite defined a project for undergraduate students enrolled in Computer Science COMP3018 called WASPS (Web Analysis and Services Pilot System). It described the situation on the ANU intranet – more than 80 web servers delivering both internal and external content, some of it indexed by external web servers but with no local search engine.

Students were asked to build a pilot system restricted to Department of Computer Science services and implementing only part of the functionality described above.

Paul's analysis of the ANU web situation

The ANU, while recognizing that Web servers and pages are established and authored freely and are thus essentially decentralised, would nevertheless like to have an automatically created and maintained, centralised database of information about these Web servers, pages and access logfiles. Web spidering software would probably help in this regard.
The database would then be used to provide additional services to the community of authors, users and/or managers of ANU's Web resources, such as:

- mapping the local Web space and access to it, and providing reports and statistical analyses of these;
- providing a feed of documents for:
  - indexing in a complete searchable document database, or archiving
  - checking for document-level properties, such as HTML compliance, broken links, documents with changed URLs, adherence to ANU link policy, adherence to a common look-and-feel policy, observance of copyright and other legal constraints, and so on.

Just before Christmas of 1998 (23 Dec), Paul submitted a *Proposal for a Programme of Research, Development and Commercialisation into the Analysis of Web Evolution: The AWE Programme*[26] which refined the ideas in WASPS and adding the idea that it should be possible to conduct searches of the ANU web as it was at some arbitrary point in time. Some students wondered whether it would be possible to search for their future results. ☺

Unfortunately, by this time Paul was already showing symptoms of the disease which would tragically kill him. One weekend, he was admitted to hospital with severe abdominal pains. Soon after that he developed balance problems and walked with a stick. By the time the fourth-year honours courses started in February, 1999 he was unable to carry out his duties. At the last minute Peter Bailey, Nick Craswell and I took over the teaching of his *Document Technologies* course. I took over as Nick's primary PhD supervisor.

Obviously Paul never had the opportunity to work on AWESOME but his ideas were the genesis of what followed in S@NITY, P@NOPTIC and Funnelback. We never did implement the "time machine" version of intranet search but we did eventually cover most of the rest. I'm keen to publicly recognize the importance of his contributions.

Our commercialisation endeavours up until the Funnelback spinoff are recounted in Chapter 3.

2.11.3 ✳TREC-7: VLC2: Even Larger Data

Paul Thistlewaite and Nick Craswell were more alive to the possibilities of the world wide web than I was, and very aware of emerging web technologies and services. After TREC-6 we set about creating a 100GB test corpus comprising only web pages. In our overview of the TREC-7 VLC Track[27] we

[26] Paul included a footnote explaining that the software developed in the project would be known as AWESOME (AWE Software Over My Enterprise.)

[27] `https://david-hawking.net/pubs/hawking_trec98vlc.pdf`

pointed out that web search engines had recently crossed the 100GB level and that Alta Vista crawled almost a terabyte.

We didn't have the ability to crawl the 20 million or so webpages which would be needed and instead asked Brewster Kahle of the Internet Archive (Wayback machine) if he would be prepared to lend us a copy of one of their web crawls. In a phone call[28] I explained to Brewster what we were trying to achieve and he agreed to supply a set of tapes.

An interesting assignment in the 1999 Document Technologies course

We asked students in the Document Technologies course to build retrieval systems capable of participating in TREC ad hoc. Most students built a component (parser, indexer, query processor) of an overall system, coded in C. One person built a parser using `flex`, another hand-coded it. Students building the indexer and query processor struggled with memory management and *segfaults*. They took a very long time and didn't get the system properly working.

As a control, one student built a complete retrieval system in `perl` which, like the original implementation of `PADDY`, used no index. To process a query it scanned all the text of all the documents, taking about an hour per query. In compensation, the time to develop the system was only about 14 hours. Amazingly, comparing the time to develop plus the time to index the data and run 50 queries, the `perl` solution easily beat the `C` competitors!

This was a very interesting assignment which taught the students a lot about Information Retrieval and about software engineering, though it was concerning that, as in most group assignments, it was hard to assign sensible marks to students working on different things.

We received a total of about 320 GB of data from the Internet Archive. Edward King of the Earth Observation Centre copied the tapes into a format we could use, and we shipped the tapes back to Brewster. Nick Craswell wrote code to filter out excessively long documents, and pages whose MIME type was other than `text/html`. He also converted the documents into TREC format, stopping once we had slightly more than 100GB of text, more than 18 million documents. In addition to the page content he included the HTTP header, within a new `DOCHDR` element. Nick organised the documents into a systematic directory structure containing compressed bundles of documents. The document IDs he assigned included the directory path and the bundlename, making it easy to extract the content of a document given its ID.

Interestingly, our choice of compression software was constrained by the deadline for shipping tapes to participants. We wanted to use `bzip2` to reduce the amount of data we had to ship, but its rate of compression was so slow that it would have taken eight days to complete the task on our high-end workstation. We used the much faster `gzip` instead.

As an aside, we found that, if documents were stored in alphabetic order of URL, the `gzip` compression ratio improved dramatically because of long strings being repeated across adjacent documents. From memory, we achieved a compression ratio of around 12 instead of the more usual 2.5.

In addition to the full corpus, initially known as VLC2 and later as WT100g, we sampled subsets of 10% (BASE10) and 1% (BASE1). The three corpora were distributed on three different tape cassette formats: DLT-4000 (two tapes), DDS-3 (four tapes), and DDS-2 (nine tapes).

Once again, TREC ad hoc topics were used. Six groups completed the full task. You can read all the details in the Track Overview.[29] Sonya Welykyj, Penny Craswell, Nick Clarke, and Angela Newey served as relevance assessors.

At the TREC-7 conference, I (on behalf of my ACSys colleagues) successfully argued for next year replacing the VLC track with a Web Track, using the VLC2 collection but using sampled web search queries instead of artificially constructed topic statements.

The story of the Web Track and ACSys/CSIRO research on web search continues in Chapter 4. In the meantime, Chapter 3 tells the very early commercialisation story.

[28] In the late 1990s, receiving a call from Australia was rare enough to get a American or British recipient's full attention. I used this to advantage on a number of occasions.

[29] `https://david-hawking.net/pubs/hawking_trec98vlc.pdf`

Chapter 3

*1999: Bringing S@NITY to ANU

Even before before Paul Thistlewaite's formal AWESOME proposal,[1] we were working flat out on implementing its ideas, and had bought a server to prototype the ANU service. In late 1998 Peter Bailey had been recruited by ANU to fulfil its requirement for in-kind contributions to ACSys. While I worked on enhancements to PADRE to support what I thought would be needed to support intranet search, Peter worked on other necessary components of the system which would launch as S@NITY on the ANU intranet in July 1999. First, he developed a parallel version of the open source `wget` web page fetcher – `pwget`.

A critical issue was to avoid exposing internal-only content via an externally-accessible search interface. Peter developed a system for testing whether URLs fetched in an internal crawl were visible externally. I believe that the server was fitted with two interfaces, one on ANU's primary B class network and the other on a C class network operated by ANU for ACTEIN, the ACT schools network. Every URL fetched via the ANU interface was tested through the ACTEIN interface. A second data set comprising only the URLs which were accessible from ACTEIN was indexed as the ANU-external collection. When requests to search the ANU-internal index were received from addresses outside the ANU network, they were silently redirected to the ANU-external index. We thought that this method would be faster than crawling once from the ANU and once from the ACTEIN interface, but it was still slow, and there were various complications. Eventually we ended up using separate crawls.

Many complications arise in crawling. A prominent one is the issue of dynamic content. It is essential to crawl URLs which include parameters because content like staff profiles may be delivered by a web script which accesses a database in order to generate the web page. E.g. `blah.anu.edu.au/staff.cgi?name=Fred.Nurke` The problem is that many webpage generators are able to generate a potentially infinite sequence of pages. A classic case is a dynamically generated calendar in which each day page has a link to `tomorrow` even if there are no events scheduled.

Another type of problem found at ANU was the spider trap where a web directory contains a link to itself. For example if `A/B` is a link to `A` and `A/page.html` is a file then the crawler will see URLs of the form `A/page.html, A/B/page.html, A/B/B/page.html, A/B/B/B/page.html, ...`. This is only one way in which duplicate or near-duplicate pages may find their way into the index. One of many other ways is that a server may have a number of aliases. For example, `www.anu.edu.au` may also be known as `anu.edu.au` and `charlotte.anu.edu.au`.

Our first attempt at avoiding duplicates was to compute a checksum and look up each new page against an index of checksums. Unfortunately this doesn't recognise as duplicates multiple copies of a page which changes every time it is visited, e.g. by including a timestamp or a visitor count. A considerable amount of work went on over subsequent years to refine tests and heuristics to reduce the amount of duplication in the index. Rampant duplication slows crawling, indexing and querying as well as increasing the size of the index and lowering the quality of search results.

URL redirections posed their own problems, "You requested this URL but you should try this one instead!" There were two main types of redirect, one generated by the HTTP server, and the other by

[1] See Page 35

the content of a web page. Redirections could form chains and sometimes even loops! It's a jungle out there on the web!

Peter Bailey developed a search interface for S@NITY. Nick Craswell worked on Perl scripts to generate administrative reports and snippets for each result as well as cached copies of search results. (Remember cached copies? Google used to provide them, but they're no longer visible for me in Australia. Nick tells me he still sees them.) One web master was able to restore his accidentally wiped out web site from S@NITY cached copies.

How to generate snippets was an important question. Older web search engines like Alta Vista merely showed the first 140 or so bytes of the document but Google had started presenting snippets containing occurrences of the query words. Anastasios Tombros and Mark Sanderson presented a paper at SIGIR 1998 in Melbourne on *Advantages of query biased summaries in information retrieval*. Initially, our system was a perl script which extracted a fixed chunk of the document text from the applicable bundle fetched by the crawler. We set Derek Foster a fourth year honours project to implement several different snippet types and conduct a user study to determine which type of summary best communicated to a human which results were likely to be relevant.

Speed of snippet generation became more and more of an issue over time. Our result pages initially included 20 results. To generate 20 snippets, the perl interpreter was started 20 times, each time uncrompressing a document bundle and extracting a snippet. Years later I brought snippet generation into PADRE, using a compact structure designed to support it and make it fast. I also later collaborated with researchers from RMIT and Microsoft on a study of fast snippet generation.[2]

1999 was in the era when web search engines started suggesting spelling corrections in the form of "Did you mean?" (DYM) links. We implemented a rudimentary DYM mechanism based on the standard Unix (English) word list `/usr/dict/words`. We needed to revisit that decision later on. See Page 138.

Tim Potter worked on packaging up all the S@NITY software in Red Hat Linux packages (`rpm`s). The idea was that S@NITY could be installed or updated by inserting a CD-ROM and using the Red Hat package manager.

In the first half of 1999, we had the S@NITY system components in place, had crawled hundreds of thousands of ANU pages and built an index. I approached Robin Erskine, Director, IT Services at ANU to inform him about the project and to prepare the ground for the launch of ANU-wide internal and external search services.[3] ANU was a co-owner of the IP in S@NITY and would be given the service for free.

At the time, ANU was running the free search engine `ht://Dig` on its main web site. Many other ANU web sites offered no search facility. It wasn't hard to argue the benefits of an effective, frequently-updated, whole-of-ANU search facility.

3.1 S@NITY Launch: 29 July 1999

We organised a significant launch event in the ground floor seminar room of the CS&IT building. Despite my evident lack of graphic design talent, I designed a brochure and had it printed.[4] I also filled a large number of display boards with posters on ACSys text retrieval research and the merits of S@NITY.

The master of ceremonies was Darrell Williamson, I gave a talk, and the service was launched by ANU's Vice-Chancellor Deane Terrell. All of us wore S@NITY T-shirts. Invitees included ACSys staff, ANU web masters and IT staff, plus representatives of CSIRO and government agencies who were considered prospects. Food was provided and the event was very well attended.

A lot of feedback was received after the launch of `search.anu.edu.au`. Dozens of web masters expressed a desire to slice and dice the search. We developed a flexible mechanism to restrict ANU search results to a single web site, a section of a web site or a group of web sites. We dis-

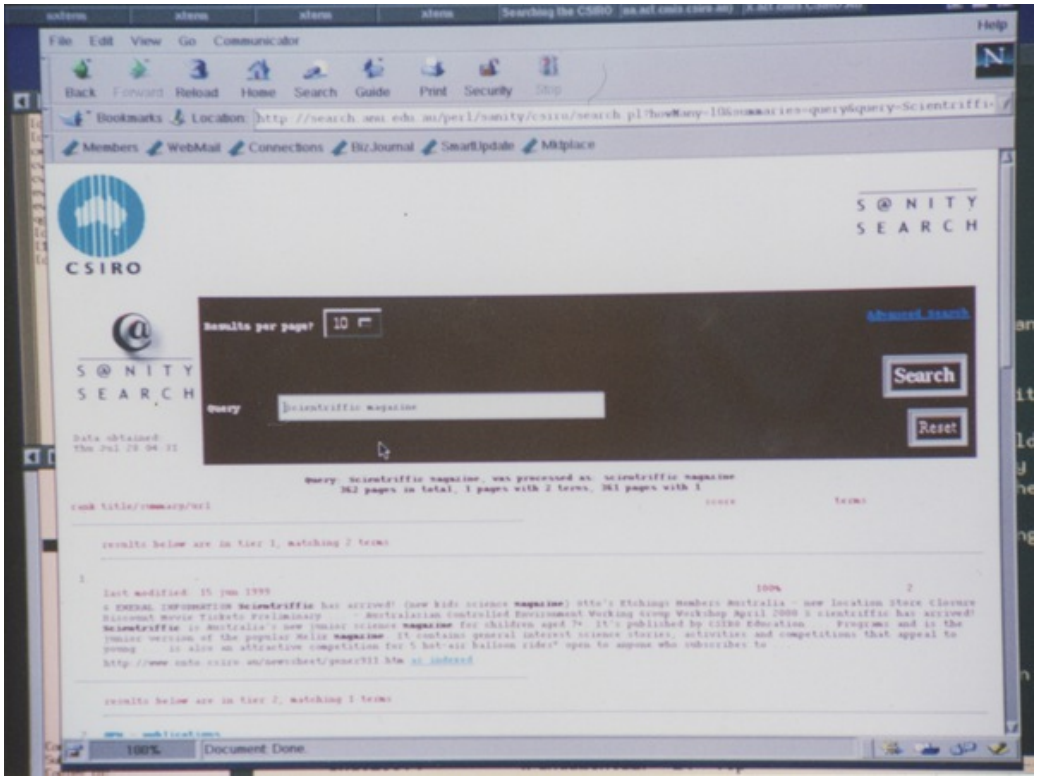
[2] `https://david-hawking.net/pubs/fp031-Turpin.pdf`

[3] See notes I prepared for my meeting with Robin in the panel on Page 41.

[4] See Appendix 11 on Page 241 for the full brochure. The rather nice S@NITY logo was created by ANU Graphic Design.

tributed HTML fragments to make it easier for web sites to provide "search this site / search all of ANU" facilities. There was a lot of discussion among customers about how broad the default search scope should be. Jakob Nielsen published advice on this in 2001[5] to the effect that one should default to broad and allow narrowing, while clearly showing what was being searched. At ANU, our advice was that the default search scope on e.g. `history.anu.edu.au` should be the whole of `history.anu.edu.au` with the ability to broaden to all of ANU or potentially to narrow to a section of the `history.anu.edu.au` domain.

29 July 1999: Launch of the S@NITY search service for ANU. Vice Chancellor Deane Terrell and ACSys Director Darrell Williamson donning S@NITY T-shirts after the opening speeches. ***Photo: ACSys***



29 July 1999: The S@NITY search interface. ***Photo: ACSys***

[5] `https://www.nngroup.com/articles/search-visible-and-simple/`

29 July 1999: At the S@NITY launch. Top: David Hawking; Middle: Roger Clarke, Chris Johnson, ANU; Bottom: Tony Boston and Monica Berko, National Library of Australia. ***Photos: ACSys***

Notes for ACSys S@NITY Proposal to ANU

Intranet Case Study – Proposal to ANU (via Robin Erskine)

1. ACSys to operate a case study on ANU intranet, including trial whole-of-ANU search service.
2. ANU to nominate a reference group with whom ACSys can consult in the development of the service.
3. ANU to allocate a machine for the purpose of running the trial.

Goals of the Case Study

1. To conduct research on intranets
2. To develop tools for managing, visualising, searching, archiving and adding value to intranets.
3. To commercialise the tools for use in other organisational intranets (IP is jointly owned by ANU and CSIRO).

SANITY – A Prototype Search Service

Search ANU's Information Treasury – Yes! — Alpha-testing started Monday, May 3, 1999

Components of the Service

1. SANITY - cgi web search interface (ACSys developed)
2. PADRE query processor (ACSys developed)
3. Document summariser - Either query-biased or not (ACSys developed)
4. PADRE indexer (ACSys developed)
5. PWGet spider (ACSys-ennhanced, parallelised version of wget)

What We Are Spidering

1. 138 servers within .anu.edu.au
2. 1329.1 MB in 168344 documents
3. Currently text/html ONLY
4. Including FTP
5. Obeying netiquette (robots.txt, delay between accesses, at night)

Security Issues

1. SANITY is only available within anu.edu.au domain so that we don't allow people outside the ANU to search (and retrieve cached copies) of pages which outsiders are not permitted to access.
2. The plan is to produce an external index by checking access rights from externally (eg. in CSIRO.)

Competitive Advantages

1. PADRE uses the most effective relevance scoring algorithm known.
2. Spidering is fast (about 12 hours) and avoids the various spider-traps etc within ANU.
3. Indexing is fast. Less than half an hour (including CRC calculations) on Peter's desktop PC.
4. Query processing is fast. A file of 120 test queries runs in under a minute. Speed depends upon query complexity. Time limits can be imposed dynamically.
5. Compared to external search engines, indexes can be far more up-to-date and more complete.
6. Duplicate detection within result lists.
7. Search on metadata as well as content, (support for Dublin Core and simple Netscape-style meta-data)
8. Automatic query weakening - Documents which match all constraints are listed first, then those which miss one etc. (Soon a MAYBE state will be introduced.)
9. Optional query-biased summaries.

3.2 Losing our S@NITY

Soon after the launch an article by Tom Worthington appeared in the Canberra Times, expressing amazement that people bright enough to create a credible search engine had naively not bothered to trademark the name. A valid criticism – I hadn't even thought of doing that. I did some hasty research and learned about trademark classes, and how to apply for a trademark.

A chain of record shops[6] was (and is still) called Sanity: [7]

It was no problem for us that they held trademarks relating to their bricks and mortar shops, but unfortunately they had recently applied to register their trademark in classes relating to online services. We would have to change our name.

After a lot of brainstorming we came up with the names "Panoptic" – seeing the whole at one view, and "Funnelback" – Funneling back important information to you, and an obvious double play on web "spidering". Both names were suggested by my wife Kathy Griffiths. We used P@NOPTIC as the name of the search engine as a whole, and Funnelback as the name of the crawler. We registered "Panoptic search" and "Funnelback" as trademarks and the `panopticsearch.com` and `funnelback.com` domains.

We soon stopped using S@NITY as the name of the ANU service, and all other pre-spinoff installations were called P@NOPTIC. If you compare the S@NITY and P@NOPTIC brochures in the appendices you will find great similarities. In the latter one I declared myself a "Panoptician" and made use of just about every word starting with "pan".

3.3 *Searching Metadata

I wanted PADRE to support library-style search as well as simple bag of words queries. Sometimes a user's information need is to find documents authored by a particular person, or published in a specified range of years. That's particularly the case when searching for publications, as in the case of library catalogue search.

Authorship and publication information is "metadata" for the document – i.e. information about the document. Unfortunately, there are many ways of representing the same category of metadata. For example, author metadata may be represented in the `From:` line of an email message, in several types of Dublin Core metadata in a web page, and in many other ways.

I assumed that it would be a useful step forward to allow searchers to specify authorship without regard to the plethora of different ways of representing it in metadata. Accordingly, PADRE98 introduced the concept of fields, represented by single lower-case letters. If you wanted to search for `brown` in the content of a document, your query term was `brown`; if you wanted to search for documents authored by `Brown` your query term was `a:Brown`, with 'a' representing all different ways of representing author metadata.

PADRE98 allowed metadata to be imported from an external metadata file. This allowed metadata from a database, e.g. a publications catalogue, to be applied to documents it referenced.

Assignment of metadata classes to field letters was done through a mapping file, configurable by searchmasters. Some letters, such as 'a' for author and 't' for title, were pre-mapped, and some fields, such as 't' were treated as part of the document content. I.e. an unfielded query term such as `explosions` would match `explosions` occurring in the title of a document as well as in its content. Weights could be assigned to matches within certain fields, enabling for example, matches within titles to be upweighted.

[6] Shops selling vinyl records, audio cassettes, music CDs, and (probably) music videos

[7] `https://sanity.com.au`

*Trials and Tribulations of Crawling

In Australia in the 1990s and early 2000s, network connections were slow, network charges were high, and web servers were typically weak. At least one Commonwealth department hosted its webserver behind a 64 Kbit/sec ISDN link! When crawling another tiny site we caused its owner to incur thousands of dollars of ISP traffic charges. The University of Western Australia blocked all accesses from ANU, in an over-reaction to our crawler having briefly overloaded a weak web server within that university.

"Netiquette" conventions suggested that a crawler should only have one outstanding request for any single server and should insert a delay between requests. Conforming to this was made difficult when single machines sometimes hosted multiple web sites and responded on multiple IP addresses.

Respecting netiquette imposed serious constraints on how quickly an intranet search engine could build an index or update it in response to changes of web content. When setting up a crawl of the Australian Bureau of Statistics (`abs.gov.au`) we were told that there were about three million web pages hosted on the site. We later learned that their servers were specified to deliver a response time of between four and ten seconds.[a] Assuming an average of seven seconds and a delay between requests of one second, a crawler could only fetch $86,400 \div 8 = 10,800$ pages per day, implying $3,000,000 \div 10,800 = 278$ days to crawl the entire site!

Crawlers exploit parallelism to speed up the fetching of content. When S@NITY launched in 1999, ANU operated about 130 web servers. A crawler running with 20 parallel threads could initially operate 20 times faster than a single-threaded one without being impolite. However, most of the servers hosted only a few pages and were very quickly mined out. Soon the crawler becomes unable to maintain 20 active threads and progress slows. Finally, we are left with one or two servers with tens of thousands of uncrawled pages. Potential solutions lie in the `sitemap.xml` convention or in differential frequency crawling based on a database of crawlng experience.

Other crawler hazards. Many organisations publish web content restricted for access by staff only. Tests based on the accessing domain name, or IP address, determine which parts of a web site are accessible. For this to work, the web master or system administrator must ensure that the webserver or firewall is correctly configured at all times. Just before a major conference organised by CSIRO and attended by many government officials, a web master within CSIRO made an error when altering a configuration file. This error opened the staff-only content to the world for a period of time during which P@NOPTIC crawler was visiting. One of the conference attendees searched CSIRO's external search service to discover some controversial internal content. That caused great embarrassment within CSIRO and was another storm for P@NOPTIC to weather.

Many web sites also include areas of content which are intended to be viewed only by humans, not by crawlers (also known as robots). Types of restricted content may include huge files, files containing data rather than text, or ephemeral dynamic content. Crawler restrictions may be encoded in a `robots.txt` file at the root of the web site, or in `robots` metatags within web pages. Differences in interpretation of these conventions led to a couple of disagreements between P@NOPTIC and web site managers.

Dangers of active URLs. Fetching some web URLs causes an action on the server, such as casting a vote. These action URLs should be protected by `robots.txt` or a similar mechanism, otherwise a crawler will initiate the action. In one amusing incident Sydney University complained that P@NOPTIC was deleting pages from its web site. Baffled for a while, we discovered that their web site administrators had published a `delete-me.cgi` script – access it with a Sydney University URL as a parameter and it would delete that page from the server. Very convenient for the administrator no doubt, when erroneous or embarrassing content made its way to the public. In most circumstances the `delete-me.cgi` URL would cause no harm, because there would be no link to it, and if there were, the crawler (be it Google, P@NOPTIC or another) would not be supplied with the parameter needed to cause damage.

However, as a public service, Sydney University published, on the web, a detailed report of web accesses, including a table of all URLs requested (as active links). In this table was included the `delete-me.cgi` URL with all the parameters which had been used by the administrators during the reporting period. P@NOPTIC fetched the access report and followed the links. Web pages which had been deleted and later reinstated were deleted again.

A similar issue was foreseen by Microsoft on its intranet. Employees like Marc Najork and Dennis Fetterly, who were researching leading edge web crawlers, were reportedly forbidden to run them on the intranet, for fear of triggering damage through active SharePoint URLs.

[a] Sites like `amazon.com` respond in fractions of a second. Amazon has reported significant revenue loss when response times extend by as little as 100 milliseconds!

The restriction to only 26 different metadata fields, and their representation by single letters rankled customers and P@NOPTIC/Funnelback implementers for more than a decade. I continued to hold the line on the grounds that it was beneficial to aggregate all the different representations of the same metadata concept, but my argument didn't apply to e-commerce sites. We increased the number of classes, and eventually, after my departure, Luke Butters modified PADRE to support strings as metadata field names rather than single letters.

We'll return to the subject of metadata later on.

3.4 *PADRE98 Query Language and Result Presentation

Given the problems ordinary people had with understanding Boolean queries, I opted for a model which I hoped would be simpler and easier to get right. At its simplest a query was just a bag of words such as `nuclear fallout shelters`. In the PADRE98 world, each of those words was a constraint to be satisfied by a retrieved document. PADRE98 actually ran an OR version of the query, i.e. `nuclear OR fallout OR shelters` but presented results in tiers, corresponding to the number of constraints satisfied. The top tier consisted of the results which satisfied all constraints (results for `nuclear AND fallout AND shelters`. The next tier contained documents satisfying all but one of the constraints, and so on. Constraints could also involve metadata fields. For example, `a:hawking web search`. Imposing date constraints was also possible via metadata fields, e.g. `a:hawking web search d>20000630`, meaning documents authored by Hawking after 30 June 2000 which contain `web` and `search`.

A number of operators were supported by the PADRE98 query language, including:

Phrase operator: double quotes, e.g. `"joe biden"` — Matches an occurrence of `joe` immediately followed by `biden`.

Alternatives: square brackets, e.g. `[dog cat "guinea pig"]` — matches `dog` or `cat` or `guinea` followed immediately by `pig`.

Mandatory presence: `+`, e.g. `+anu` — `anu` must be present in every returned document.

Mandatory absence: `-`, e.g. `-anu` — No returned document may contain the word `anu`.

Tilde: e.g. `~anu` — The word `anu` is not treated as a constraint, but the relevance scores of documents containing it are boosted.

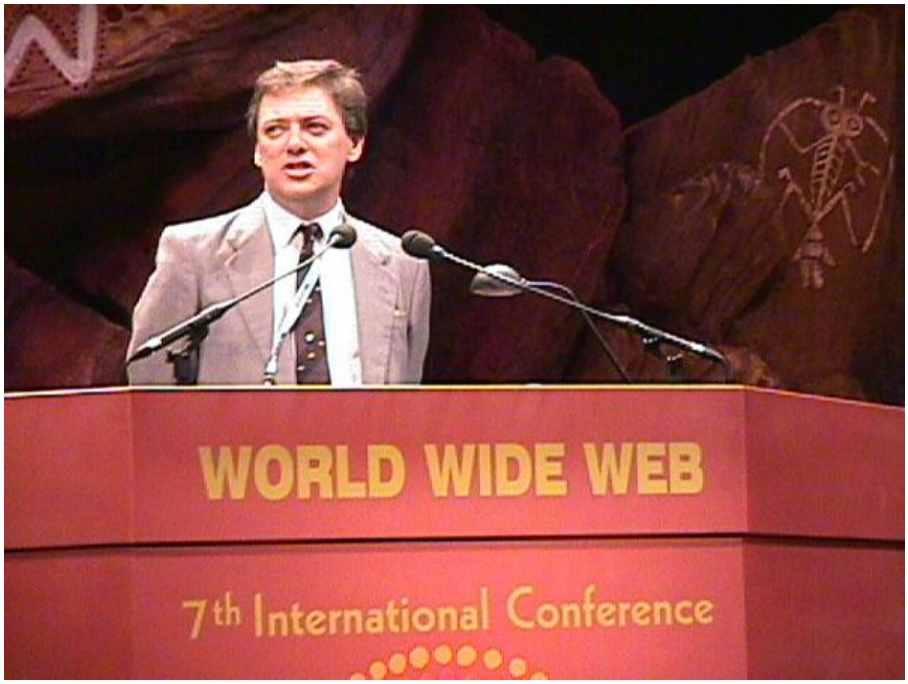
As I write, this model still makes great sense to me, but many customers didn't like it because it wasn't how other search engines worked. Regrettably, neither P@NOPTIC nor Funnelback gained enough traction to persuade customers to follow a different path.

I envisaged that queries involving metadata constraints would often be generated through a web form, or inserted behind the scenes, rather than written by users. However, many people don't realise that typing a query with operators allows for much more flexibility than a form does. I've seen organisations deploy search forms which were semantically ambiguous, or are too complex to enable users to work out their meaning, ... while still being unable to match the power of a text query. Indeed, such web forms are quite common.

3.5 *Web Search Versus Text Retrieval

A problem which soon became apparent with S@NITY/P@NOPTIC was that, although we were providing state-of-the-art *text retrieval* we weren't providing state-of-the-art *web search*. The same was the case for most of the pre-Google web search engines. If you searched for `white house` on one of those engines, the home page of the White House would be buried deep among the results. Reviewers like *Search Engine Watch* lambasted web search engines for not being able to find their own company homepage!

In 1998, as we began thinking of commercialisation, the world started to become aware of Google. That year at the WWW7 conference in Brisbane, chaired by Paul Thistlewaite and attended by Nick Craswell, Sergei Brin and Larry Page presented their paper on Google, *The anatomy of a large-scale hypertextual web search engine*.[8] Their Google prototype[9] was hosted at Stanford University and indexed 24 million web pages.[10] Google also submitted a paper on PageRank to the SIGIR 1998 conference but it was rejected.



Program Chair Paul Thistlewaite addressing the audience at WWW7 in Brisbane. ***Photo: WWW7 organisers.***

PageRank was a way of assigning importance weights based on hyperlink structures of the web. A web page is determined by the number of incoming links and the PageRanks of the referring pages. A more authoritative page will have a higher PageRank and its outgoing links will be weighted more heavily. An iterative computation eventually converges to a set of PageRank scores. Scores lie in the range 0–1 and can be thought of as the probability that a random web surfer would land on that page.

JumpStation – The first Web search engine – December 1993 (and a knife fight in a pub)

JumpStation,[a] which was built by Jonathan Fletcher, a systems administrator at the University of Stirling in Scotland, is now generally regarded as the first web search engine which looked and worked the way current search engines do. It was tiny and folded a year later due to absence of investment.
Jonathon Fletcher was honoured at the SIGIR 2013 conference in Dublin, Ireland. He and I were chatting while lining up for a Guinness in a pub on the Liffey when a couple of men started fighting. A cry of, "He's stabbing me!" was ignored by all the locals until blood became obvious and a small pen knife clattered to the floor near me. I stood on it. A barman threw the two men out on the street and seemed to think that that was, "problem solved."

[a] `https://en.wikipedia.org/wiki/JumpStation`

A different way of using link information was embodied in IBM's `Clever` search engine which implemented Kleinberg's HITS algorithm.[11] HITS distinguished between hub pages and authoritative pages. The algorithm assumed that good hubs link to good authorities and good authorities are linked to by good hubs. It used an iterative algorithm to assign hub and authority scores.

[8] `http://infolab.stanford.edu/~backrub/google.html`
[9] Google evolved from BackRub.
[10] That's only about one-third larger than the VLC2 collection which we started shipping in June that year.
[11] `https://www.cs.cornell.edu/home/kleinber/auth.pdf`

An underlying assumption of link-based methods is that web pages are authored by humans who only link to pages they wish to recommend. That assumption doesn't always hold but link-based methods are useful in identifying web pages which users like to see in results lists, *when they are relevant*.

Encouraged by Google marketing, the media generally assumed that the fact that Google worked well was due to PageRank.[12] However, the Google Anatomy paper makes it clear that use of "anchor text" is critical to the identification of important answers to a query. Anchor text is the text that people click on in a web page to follow a link. Sometimes the concept is extended to include the enclosing sentence or paragraph.

Google indexed the anchor text with the target as well as the source of the link. To illustrate how this works, consider the anchor text, `Australian National University`. Across the web there may be thousands of links which use this exact anchor text and nearly all of them reference `www.anu.edu.au`. If `Australian National University` is received as a query, there is compelling evidence that `www.anu.edu.au` is the best answer. Anchor text also permits retrieval of a page for queries whose words are not present in the page. E.g. retrieving images, or responding to queries in foreign languages or which contain common misspellings. In the early days Google used anchor text to return links to pages whose URLs it was aware of but had not crawled.

People who did understand Google's reliance on anchor text started the game of "Google bombing." They created links from a large number of pages, all using identical anchor text. For a while this led to George W. Bush's home page being the top ranked result for `miserable failure`.

In addition to PageRank, Google's Anatomy paper lists a lot of static features which should be taken into account when ranking:

> ... reputation of the source, update frequency, quality, popularity or usage, and citations.

In 2000 Larry Page said that the Google ranking function included more than 40 variables. He intially set the coefficients manually and said that when Monika Henzinger (at the time Google's Chief Scientist) later did an exhaustive tuning exercise, she found that only small changes were needed to the coefficients. The Anatomy paper mentions that the Google ranking function included a feedback mechanism, presumably based on user interaction data.

I believe that the Google ranking function became more complicated later but although my PhD student Trystan Upstill joined the Google search quality team in 2005, and later led it, I could never persuade him to spill the beans. Apparently new recruits are trained to resist both the wiles of former PhD supervisors, and blow torches.

Web search engine result ranking functions combine static (query independent) and dynamic scores. Dynamic scores are query dependent and derived from the document and its referring anchor text. The dynamic part of Google's ranking function took account of proximity between query word occurrences and up-weighted occurrences in boldface, capitals or larger font.

In a section on design goals the Anatomy paper criticises other search engines for being secretive and advertising-oriented. Rather ironically, they said:

> With Google, we have a strong goal to push more development and understanding into the academic realm.

From a user point of view, early Google was a major step forward: it responded very quickly indeed; it was free of poorly targeted advertising; it generated query-biased snippets; it seemed less prone to spam; and, above all, it did a much better job of satisfying the intent behind common user queries.

At the time, Brin and Page were still PhD students at Stanford but their paper acknowledges funders and names three computer companies who had donated equipment.

In around 2000, Andrei Broder, then Chief Scientist at Alta Vista, proposed a taxonomy of web search, classifying queries into the following categories:

[12] Google is good. Google uses PageRank. Therefore, PageRank is good. Jaguar cars have leather seats. Jaguar cars go fast. Therefore, leather seats make cars go fast.

Informational – e.g. `ferry sinkings`.
Transactional – e.g. `buy Sony VAIO Z laptop`
Navigational – e.g. `Australian National University`

He also spoke of the importance of satisfying the intent of a query rather than finding documents which are relevant to it. I.e. *utility* rather than *relevance*.

Concerned that TREC was too focused on newswire items, I persuaded Donna Harman and Ellen Voorhees to invite Andrei Broder to speak at TREC-9 in 2000. This was against convention, since only participants and sponsors had ever been invited. His talk was rapturously received by the audience.

3.6 *S@NITY Versus Web Search

When launched in 1999, S@NITY was a web search engine operating on an intranet (a small web), but using a ranking function designed for text retrieval.

It soon became apparent that S@NITY, as launched, was not capable of handling navigational queries. Bob Williamson loudly complained that S@NITY didn't give appropriate responses to the query `IP policy` and that he used Google instead to search ANU.[13]

We needed to urgently start using anchor text. Starting with his internship at Microsoft Research, Nick Craswell led a research study into using anchor text for site finding[14] which dramatically confirmed its value. Nick's anchor text system built surrogates for actual pages by concatenating all the referring anchor text and then did standard text retrieval over the surrogate collection.

I worked on building the necessary structures to support use of anchor text in combination with document content and metadata in PADRE. Referring anchor text was indexed as 'k' metadata.[15] Once this change was implemented, PADRE allowed documents to be ranked taking into account scores derived from anchor text, titles and page content with tunable weightings.

The "poster child" example of poor ranking at ANU was the query `library`. Ranked using our state-of-the-art text ranking algorithm, the ANU library home page ranked about 100,000 places down. Ranked first was *Biology cancellations for 1996*. It contained dozens of occurrences of the word library in a relatively short document and thus achieved a very high BM25 score. By contrast, `library.anu.edu.au` contained only one occurrence.

Adding anchor text to the mix promoted `library.anu.edu.au` to rank one, followed by the library catalogue and homepages of various subsidiary ANU libraries. — Nailed it! ☺

Aside: Nick Craswell points out that we may have been able to spot ranking problems without relying on user feedback, if we had scaneed the logs for common query reformulations such as: `library` → `anu library` → `anu library home`.

Later on, when building a demonstration search for Curtin University, we found that using BM25 with standard parameters wasn't the optimal way to score anchor text. Curtin university contained a number of libraries and our then out-of-box ranking function failed to return them in the preferred order for the query `library`. Study of the situation revealed that the default BM25 settings caused the scores of documents to be length normalised. Normalisation largely nullified the fact that the main library received much more anchor text than the subsidiary libraries. We feared that BM25's treatment of *term saturation* might also be problematic. Accordingly, Trystan, Nick and I devised a new function called AF1 (Anchor Formula 1) and used it to fix the demo. We disclosed it as a poster at SIGIR 2004,[16] under the title *Toward better weighting of anchors*.[17]

Trystan Upstill remembers this as an example of how our small group did science:

[13] Even though that missed out on internal content.
[14] `https://david-hawking.net/pubs/craswell_sigir01.pdf`
[15] See Page 42.
[16] `https://david-hawking.net/pubs/hawking_sigirposter04.pdf`
[17] Our practice toward protecting IP was to publish it, and thereby prevent anyone from patenting the idea against us.

I remember on many occasions swinging past your office with a vague idea we'd been discussing, and talking it through, then jamming with `gnuplot` to play with the estimations. A clear one that comes to mind is when we were playing with anchor credit. We'd been thinking about a bunch of capping strategies – Nick's work applied a BM25 squash but that was clearly hurting in this case. Thinking about it more generally, it seemed that for anchor text it may be that "the longer the better" applies after all! Then we realised we'd just been thinking too hard: why not just add it on a log scale! `<boom>`

I loved the simple pragmatic way in which everyone I worked with operated. I recall many months of work trying to figure out why we couldn't show retrieval benefit from use of PageRank. Then there was the Kraaij, Westerveld, and Hiemstra paper[18] that showed a basic (grep-like) URL classification performed better in homepage finding. After we chatted about it we realised that was too complicated too – simple URL length gets you 9/10 of the way there in terms of understanding importance. And it has the added bonus that it's darn cheap to compute! This Occam's-razor approach informed pretty much every piece of engineering I've done since: it's not always that deep. ☺

I remember every time we interacted felt super dynamic, it was like a free-form explorative chat which we could then go and play with. Working with the team gave me a unique opportunity to experience experimentation on a production scale in a world leading capacity. To this day I feel incredibly grateful I was given this opportunity to research, understand and play in such a supportive environment.

Attending conferences with Panopticians was always great,[19] everyone on the team was super chill, but also respected as people who didn't just Talk the Talk, they Walked the Walk. As a junior researcher this meant I was able to feel like I was in the absolute thick of the research field, but also building a highly effective product that impacted people's lives (and that I used daily!).

I have so many memories of sitting outside the Purple Pickle,[20] I forget what I used to drink, but I do remember the affogato sitting at the table somewhere.[21] Sitting in the sunshine discussing everything from the minutiae of ranking formulations or index construction, or trying to solve the world's problems is a memory I hold particularly dearly.

Trystan also remembers a couple of my mantras:

- *Never travel without a Frisbee!* He recalls playing Ultimate with me on the Microsoft Research campus at J.J. Thomson Ave in Cambridge, UK. That reminds me of throwing a Frisbee with Steve Blackburn in Amherst MA, while standing in ankle deep snow, and of a very successful Ultimate match held at SIGIR 2002 in Tampere, Finland. Dozens of serious academics joined in!
- *Never hide your light under a bushel!* This is not a natural part of my personality, but something I learned as an essential aid to survival and success in academic research. It was encouraged and expected by grant agencies, promotion committees, annual performance appraisals, and project reviews. I'm confident I never extended the mantra as far as artificially magnifying my own light, or suppressing the light generated by others.

At SIGIR 2004 in Sheffield, Amit Singhal, then head of Search Quality at Google, was sitting in the poster area and asked me if we had a poster. I said yes, and he started "Googling" for it. He couldn't find it because it wasn't yet on the web, and demanded to know how I expected anyone to read my work if it wasn't online before the conference. I suggested that he stand up and walk five metres to the relevant poster board. ☺

Although ANU's intranet was a microcosm of the web, there were several important differences. Distinction between internal and external content was one. Another was that in an intranet there is almost no spam publishing because there is no incentive for it – there were no ANU e-commerce sites. Other differences include the restricted scale of ANU's intranet, and the relatively low volume of queries submitted to its search service – initially of the order of a thousand queries a day to each of the internal and external services. Low query rates reduced pressure on the search infrastructure but also limited the degree to which user interaction data could be used in ranking or AB testing.

[18] `https://dl.acm.org/doi/abs/10.1145/564376.564383`

[19] Trystan's experiences at SIGIR weren't always so great. After the SIGIR 2007 banquet in Amsterdam, he was mugged while carrying home a silver most-connected-author trophy which he'd accepted on behalf of a Google colleague.

[20] The coffee shop across from the CSIT building at ANU, also known as "Meeting Room PP."

[21] Probably mine.

ANU's staff were not backward in complaining about perceived shortcomings in the search service. One wrote an angry letter to the Vice-Chancellor complaining that information about a grant scheme could not be found via S@NITY. The complaint found its way to me and I replied that,

> All search engines are essentially statistical in nature and can't be guaranteed to work perfectly in every case. However, we are scientists and we are very keen to improve our algorithms. Would you kindly tell us what you searched for (the query) and the page(s) you think that S@NITY should have found.

A couple of days later came the reply,

> Oh dear. I appear to have done you a disservice – the information I needed was on the Australian Research Council web site, not within ANU.

Many cases like this arose, and many other challenges raised their heads as we attempted to use P@NOPTIC to meet our CSIRO external earnings targets. ACSys came to an end in September 2000, leaving both CSIRO and ANU with rights to independently pursue small-scale commercialisation of the technology. All subsequent commercialisation was in fact undertaken through CSIRO. On the ANU side, Paul Thistlewaite had died, I had joined CSIRO and Peter Bailey had also left. Chapter 5 tells the story of commercialisation within CSIRO. The timeline of that chapter overlaps that of the research program we were conducting in parallel, described in Chapter 4.

Chapter 4

*1999–2008: CSIRO Research on Text Retrieval and Web Search

While development and commercialisation of P@NOPTIC continued apace, scientists and PhD students in and around CSIRO continued to strive for an understanding of all aspects of search. Initially, much of our work tried to make use of the collaboration possible under the TREC umbrella, through the Web Track. Over the six early editions of the TREC Web Track (1999 – 2004), participant understanding of web search grew, and an evolving series of research questions were addressed. I was Track Co-ordinator or co-Co-ordinator of each of these six editions.

`es.csiro.au`

Constant organisational change within CSIRO made it difficult to maintain consistent branding and a stable web presence, while trying to build an international reputation. Some of the early electronic resources of the Technologies for Electronic Documents group, including publications and data sets were hosted on `pigfish.vic.csiro.au` but the group was reorganised and the leased `pigfish` hardware was returned to Dell before its contents were saved. My defence against a repetition of this small disaster was to create, in 2002, a web site called `es.csiro.au` and host it on a purchased machine rather than a leased one. According to the WayBack Machine the site was still active up until December 2017.
ES stood for Enterprise Search, but it was deliberately not the name of any group/project/team/stream. To encourage people to cite our research, full-text author versions of every one of our publications were hosted on `es.csiro.au`.[a] Nick Craswell used his perl and web expertise to generate the publications page from the BiBTeX file which we had to produce anyway. With his permission, my personal web site `https://david-hawking.net` still uses the same marvellous technology.

[a] This horrified many of our fellow CSIRO researchers but we caught the wave of open publishing. We only published in the rapidly increasing number of venues which permitted this. Many academics wanted everyone to be able to access their publications free of charge, and felt very strongly about it. In 2001, 40 members of the editorial board of the Machine Learning Journal resigned in protest at lack of free access to its articles.

I apologise if this chapter seems disorganised. A section describing what was done and what was learned in each of the Web Track editions appears in temporal order, but the sequence is occasionally broken by sections describing work arising in the context of the Web Track but pursued outside it. The sequence of sections describing the Web Tracks is followed by a section on the TREC Enterprise Search Tracks 2005–2008. A final section groups together search research conducted by CSIRO and students outside the context of TREC.

4.1 ∗1999: The TREC-8 Web Track

Two tasks were offered in the TREC-8 Web Track:[1] the Large Web Task (using 100 gigabytes of web data and queries from web search engine logs), and the Small Web Task (using 2 gigabytes of web data and the TREC-8 ad hoc topics.)

Section 2.11.3 on Page 35 describes how we created a 100GB corpus. It was originally called VLC2, but later known as WT100g. Many academic organisations found it impossible to work with a corpus of that size. Accordingly, we heuristically created a 2 gigabyte subset called WT2g. From the track overview it seems that we[2] selected hosts with highest relevant density from the previous year's tasks and included all the pages from each selected host. The include-whole-hosts strategy gave WT2g better connectivity because within-host links are common.

Nick Craswell created a hyperlink connectivity graph for WT2g and ACSys made it available to participants.

Small Web Task: We found a high and significant correlation between retrieval performance on web data and retrieval performance on the ad hoc task, using the same topics, and binary judgments from NIST assessors.

In the Small Web Task, no measurable benefit was gained from the use of link information in ranking. This result triggered calls of derision from TREC attendees, who "knew" that Google results were so good because of their use of link information. In our overview paper, we raised three questions to try to explain this observation:

1. Were there enough cross-server links in WT2g?
2. Would the advantage of links show up in types of information need such as homepage finding?
3. Would link-based methods come to the fore if judging were multi-level rather than binary with a weak threshold? (I.e. topic distillation. See the panel below.)

None of the participants appear to have used anchor text in ranking.

What is Topic Distillation?

While working at Alta Vista in the late 1990s, Krishna Bharat proposed the concept of *topic distillation* in web search. This encapsulated the idea that if you were searching for e.g. `Sony SLR cameras` there are likely to be a small set of resources for that topic which could satisfy your need. For example, Sony's web page on SLR cameras, a Wikipedia article explaining how SLR cameras work, reviews of Sony SLR cameras, and pages which allow you to buy Sony SLRs.

Large Web Task: In this task, the queries were no longer NIST-constructed topics, but rather 10,000 "natural language queries" sampled from Alta Vista (thanks to Monika Henzinger) and Electric Monk (thanks to Edwin Cooper) query logs. Alta Vista provided an interface by which queries expressed as sentences or questions could be submitted. Examples included: `why is the sky blue?`, and `how can I order flowers online?`.

Electric Monk was a search engine oriented entirely toward natural language queries. A front end developed by Edwin Cooper, son of the well-known information retrieval researcher Bill Cooper, subjected the queries to syntactic analysis, leading to a Boolean query which was submitted to Alta Vista. It would have been interesting to empirically evaluate whether the Electric Monk improved on Alta Vista results, but we were focused on the Web Track.

We chose natural language queries because we thought that it would be more likely that judges could decide on the intent behind the query. As you would expect, the search engine logs included a high proportion of "adult" queries. Some participants requested that those queries be filtered out prior to selecting the 10,000 queries to be processed. I descended to the limits of my depravity in composing a black list of words and phrases indicating adult content. I would run the script, scan

[1] See the track overview at `https://david-hawking.net/pubs/hawking_trec99wt.pdf`

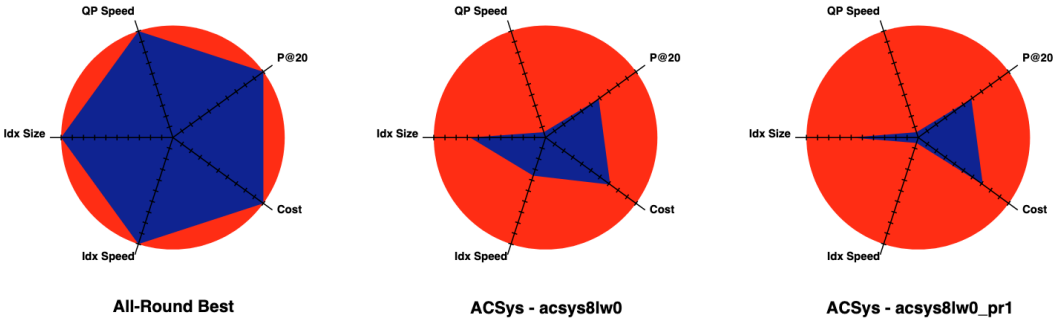
[2] Nick Craswell did the actual work!

the output, add more strings to the black list and then repeat, until I had achieved the required level of filtering. It wasn't easy – Does `black escorts` refer to Ford motorcars or to a sexual service?

To avoid participants totally ignoring efficiency or focusing their system on particular queries, Web Track participants were required to run all 10,000 queries and send the top 20 results for each to NIST. We, the track organisers, later randomly selected 50 queries for assessment by ACSys judges. Judgments were still binary, and still focused on relevance rather than utility. There was no penalty for returning duplicate documents.

Systems were compared on five dimensions: cost, indexing time, index size, query processing time, and precision at 20 documents retrieved (P@20). Results were presented in the form of Kiviat diagrams. (See the example on Page 52.)

Two participants also reported scale-up data for their systems, having performed runs on the 1% (BASE1) and 10% (BASE10) samples of WT100g. Nick Craswell computed PageRanks for the pages in WT100g and unsuccessfully attempted to use them to boost the quality of result rankings.



1999: In the TREC-8 Large Web Task, Kiviat diagrams were used to compare performance on five dimensions. Measurements for a system on each dimension are linearly scaled relative to the the best performance achieved. The plot on the left shows the performance of a hypothetical system which achieved best scores on every dimension.

4.1.1 *Showing Off Again Leads to Unintended Hilarity

At the TREC-8 conference I again attempted to show off by demonstrating that the Large Web task could be completed on a mid-range laptop. As my presentation of the Web Track overview began, I started running the Large Web queries against a PADRE index of the full WT100g on my laptop. I projected the results on one screen and my talk on another.

As I was explaining to the audience the difficulty I faced in eliminating every last adult content query from the set of 10,000, there was a gale of laughter from the audience – far louder than anything I had ever managed to achieve with intentional jokes. I looked up at the screen running the queries and discovered that the last query processed had been `blow job` – apparently I had filtered out the plural but not the singular. ☺

4.2 *Web Size and Search Engine Coverage

In the late 1990s there was considerable scientific interest in measuring the size of the indexable web, and how much of it was accessible through web search engines. A 1999 Giles and Lawrence paper published in Science Magazine, showed relatively poor coverage by search engines, and little overlap in their content. A 2000 paper (*Chart of Darkness*[3]) by Peter Bailey, Nick Craswell and me used privileged access to web sites within ANU to demonstrate that a substantial proportion of ANU web content was not able to be crawled either by an internal ANU search engine or by external web search engines.

[3] `http://david-hawking.net/pubs/bailey_tr00.pdf`

Lack of overlap in search engine coverage was used as an argument in favour of metasearch[4] – it was argued that by combining results from multiple search engines with only partially overlapping coverage, a more comprehensive, and therefore higher quality, results list could be produced.

Around that time, there were a number of companies promoting themselves on the basis of determining ever larger percentages of the web which were not being crawled and indexed. Some claimed to have the ability to index *the dark web*. In fact, there would be no point in indexing most of this so-called *dark matter* since a lot of it was not usefully searchable data, such as personalised content, ocean temperatures and telemetry from space probes and satellites. If you allow for dynamic content generators, like calendar pages, the size of the web is clearly infinite.

The idea that the size of a search engine's index (beyond a certain size) was critical to the overall quality of search results was soon debunked. In the following section, we show that in 2000/2001, although Northern Light's index was many times larger than Google's, Google outperformed Northern Light. In the case of navigational searches, Google achieved a success rate around five times higher than Northern Light. MetaCrawler, the dominant metasearcher at the time, ranked no better than fifth in our rankings, despite its theoretically massive coverage. However, the experimental metasearcher Inquirus, developed by Steve Lawrence (an Australian working in the USA), performed well on the informational task.

4.3 *Measuring Search Engine Quality

In the TREC-8 Web Track, ACSys assessors judged the relevance of documents retrieved for "natural language queries" sampled from logs supplied by Alta Vista and Electric Monk. Since the WT100g (aka VLC2) collection was actually a small and out-of-date web crawl, we realised that the systems used by TREC VLC participants could be regarded as web search engines, and their performance compared with that of real public web search engines. The real engines would be expected to operate at an advantage because of their larger and more recent index coverage.

We fed the queries used in TREC-8 VLC to twenty search engines, retrieved the resulting pages, and employed the same ACSys assessors to judge their relevance. To our surprise the TREC VLC participants performed very creditably in comparison to the real search engines. See the bottom graph on Page 54.

We published an article on *Measuring search engine quality*,[5] which was, for quite a while, one of the most frequently downloaded articles from the Journal of Information Retrieval.

4.3.1 *2000: The Infonortics Search Engines Meeting

David Evans was a TREC-8 participant and adviser to Everett Brenner, the Program Chair for the 2000 Infonortics Search Engines Meeting in Boston, MA. I was invited to present our results at that meeting and David explained that he was going to use me as a "gadfly" to provoke the web search engines into action on improving search quality. He would position my talk just before a panel session involving the leaders of several of the search engine companies.[6]

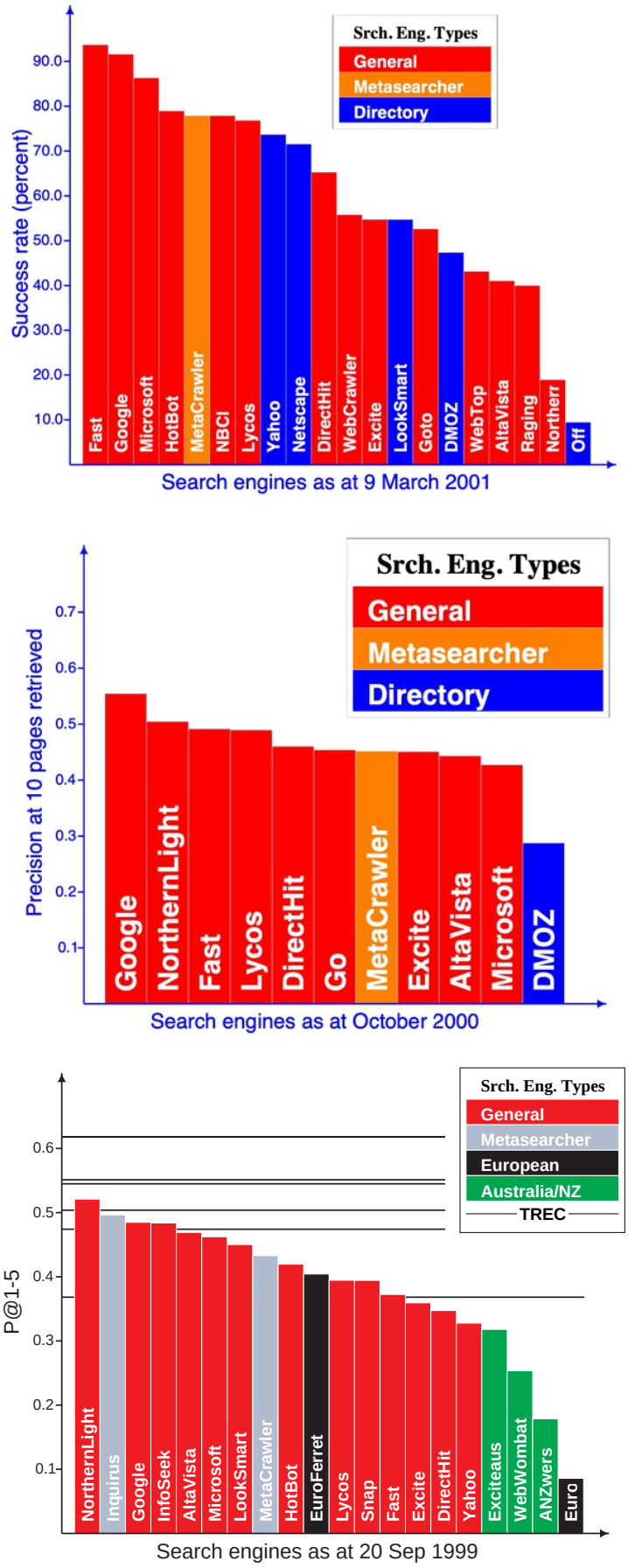
Unfortunately, I broke my fibula while playing ultimate frisbee just prior to flying to Boston, and travelled with a splint on my lower leg and a crutch. For a moment I thought that my injury would give me a good case to sit in an exit row on the plane, but Qantas told me that the opposite was the case – injured people are forbidden to travel in exit rows.

At lunch before my talk, David Evans marshalled me and some of the panellists onto the same table at lunch, and told the search engine people about our evaluation work. I was seated next to Andrei Broder and he asked me a few questions before expostulating, "binary judgments!" and turning his back on me. Rather than argue with him, I decided to wait for my presentation.

[4] A metasearcher is a service which aggregates the results from multiple search engines.

[5] https://david-hawking.net/pubs/hawking_ir01.pdf

[6] The programme for the 2000 Search Engines Meeting is archived at https://web.archive.org/web/20000511130125/http://www.infonortics.com/searchengines/boston2000pro.html



2000 – 2001: Our evaluations of the quality of web search engines on each of Broder's information need types: Top: Navigational, Middle: Transactional, Bottom: Informational – notice the very creditable performance (horizontal lines) of participants in the TREC-8 Web Track.

2002: Ev Brenner, Program Chair, Infonortics Search Engines Meeting

I argued the case for objective evaluation, along TREC lines, and presented the results of our evaluation. I aimed for a bit of theatrics and remember waving my crutch at the audience to emphasise a point. When the panel started, I was invited to sit across from the panellists and respond to their positions.

It was a star-studded panel, comprising Larry Page (co-founder of Google), Eric Brewer (co-founder of Inktomi, which powered several web search engines including Yahoo!, Microsoft, and HotBot), Jan Pedersen (Chief Scientist at Infoseek), Marc Krellenstein (CEO of Northern Light), and Andrei Broder (Chief Scientist at Alta Vista). Larry Page attacked the need for objective evaluation. He said, "At Stanford we built a bunch of stuff, and you just know what works and what doesn't." Later, "In my drawer I have an evaluation which I wish I could share with you. We got a company to do a survey of Google users and found that a huge percentage were very satisfied. That's the only evaluation we need." I responded by asking him whether that meant that Google no longer needed to do R&D aimed at improving their system. (Since without careful evaluation you can't really know whether you are improving.)

After the day's sessions, Larry Page sought me out in the hotel lobby, and apologised for attacking my position, explaining that he thought that there was no point in having a panel if there was no controversy or debate.

In the 2000 meeting other key search engine people were also present including Knut Magne Risvik, Chief Scientist of FAST Search and Transfer / All The Web. I attended several subsequent Infonortics meetings, presenting and chairing sessions, but the web search companies no longer bothered attending.

As you can see in the graphs on Page 54, we extended our evaluations to include Broder's other information need categories, Transactional and Navigational. In the Navigational case, major differences appeared between those search engines which were using web evidence, like referring anchor text, and those which were not.

Looking back, Andrei Broder's scorn about the use of binary judgments was largely justified in the case of Informational and Transactional information needs, but it depends upon where the cutoff is drawn. The critical question is whether a document satisfies the information need behind the query. In TREC ad hoc, the implicit information need is to find *all* the documents which relate to the topic. Maybe binary judgments are fine in that case, but in web search, information needs are extremely rarely like that. Thinking of retrieval as topic distillation, the number of documents which singly or together satisfy the user's information need, is usually very small. It may be OK to use binary judgments in that case, with a very stringent criterion: Either a candidate is in the topic distillation set, or it isn't. However, it's better to allow grades of usefulness to recognise different

levels at which documents are useful in solving the information need. Multi-level judgments are also useful in Learning to Rank.

The judging situation becomes much more complex when you start thinking, as Tim Jones and I did, about *conditional utility*,[7] where the utility value of a result is not decided in a vacuum but depends upon what results have already been presented.

In Navigational search, and in known-item search, binary judgments are generally quite adequate.

4.3.2 Subsequent Infonortics Search Engine Meetings

Infonortics was a private company run by a Harry Collier, an Englishman with a taste for luxury hotels, and an abhorrence of fuss, bother, and delay. Registration fees were high, but Harry waived the registration and a couple of nights accommodation at the conference hotel if you were a speaker. I attended several Search Engine Meetings, contributing as a speaker and as a session chair.

Before one Search Engines Meeting, Harry expressed exasperation at delays frequently caused by speakers insisting on using their own laptops and having difficulties getting the presentation working. He insisted that all presentations be emailed to him in advance for loading on his computer. When Harry's video card blew up over lunch I, as chair of the next session, was rather dismayed. One of my four presenters hadn't brought a computer or any copy of his presentation.

After re-ordering the running order for the session, one speaker with their presentation on a laptop kicked off the session while another two took a laptop up to a hotel room and remotely logged in to a university network to locate and retrieve their presentation. The fourth presenter had a thumb-drive and used the first speaker's laptop. The session finished on time and without any significant change-over delays!

On my way to the event in San Francisco, I ruptured an eardrum descending from altitude – sudden relief from pain followed by an ability to blow air out of my ear. On landing I shared a van with several other people. All went well until the driver was pulled over and given an infringement notice by a highway patrol. This was apparently the second such ticket that day for our driver. He totally lost his cool, claiming that the police had followed him from the other side of the city and were bent on persecuting him. After the police left, he seemed to enter a psychotic state, shouting at the top of his voice:

– Do you know who I am????
– Do you know who I AM!!!!
– I know George Tenet!![8]
– Do you know who I AM?!!!!!!!!
– I came from the River Niger to save the people of America!!!

Several passengers found that they urgently needed to get off earlier than originally planned but I noticed that we were stopping at red lights and yield signs, correctly indicating and staying in our lane. I stayed on until my destination.

4.4 *2000: The TREC-9 Web Track

There were again two web tasks in TREC-9, the Main Web Task and the Large Web Task.

Main Web Task. Realising the limitations of WT2g as a useful corpus for investigating web search, ACSys set about engineering a more suitable, better connected corpus. The process is described in a paper by Peter Bailey, Nick Craswell, and me, *Engineering a multi-purpose test collection for web retrieval experiments*,[9] and the result was called WT10g. Nick Craswell again computed a link connectivity matrix and it was distributed with WT10g.

[7] See panel on Page 66.
[8] CIA director, 1997–2004.
[9] https://david-hawking.net/pubs/bailey_ipm03.pdf

Test topics were engineered by NIST assessors from an eXcite query log (provided by Jack Liangjie Xu, who was a TREC participant through Berkeley University as well as a leader at eXcite. The title field of a topic was the unedited query, while the NIST assessor made up description and narrative fields to describe an information need they imagined might lie behind the query. Several of the title fields contained apparent spelling errors. Judgments were made relative to the full query. This meant that valuable answers to an alternative interpretation of the query were judged irrelevant.

In a first for NIST, judgments were ternary (irrelevant, relevant, highly relevant) and judges were asked to identify the 'best' answer(s) for each topic. Also for the first time, Normalised Discounted Cumulative Gain (NDCG)[10] scores were calculated.

This could have been a topic distillation task but unfortunately, at least some of the assessors were not web-savvy. For one topic about "activities in Californian national parks", a perfect topic distillation answer comprising lists of links to parks, recreational activities, accommodation, entry fees, supplies, and gear, was not regarded as highly relevant while some of the highly relevant pages were of no real use to someone planning activities in Californian national parks.

Participation categories typical of TREC ad hoc were defined: automatic v. manual, title-only v. full-topic.

Although some participants had started using referring anchor text, link methods failed to bring any benefit to the main informational task. In contrast, ACSys separately showed that the links in WT10g were sufficient to bring a dramatic benefit to a homepage finding task over the same corpus.[11]

Large Web Task. This was very similar to the TREC-8 Large Web Task, but after runs of the 10,000 queries had been submitted a decision was made to investigate service finding (Transactional) rather than pure Informational retrieval. 106 queries which seemed to be designed to locate online services were selected, and the ACSys judges were asked to judge whether submitted URLs provided access to the requested service or not. An example query was `where can i find an online translator?` For the first time judges viewed candidate documents using a browser, albeit a text-only one (`lynx`). The use of hyperlink and anchor text information was found to bring very little benefit on this task.

4.5 *2001: P@NOPTIC Expert

As noted on Page 47, Nick Craswell's work on retrieval via anchor text involved creating surrogate documents and running the retrieval system over the surrogates. The surrogate for a web page like `whitehouse.gov` was a concatenation of the anchor text from every incoming link to that page. When processing a query like `george w bush`, the retrieval system would find many surrogate documents matching the query, but the surrogate containing the highest density of matches would rank top – i.e. the one with the largest number of relevant referring anchor texts.

Nick had the idea to extend this idea to expertise finding. Starting with a crawl of the CMIS (CSIRO Mathematical and Information Sciences) web presence and a staff list, he used a `perl` script to extract those passages of text from the crawled web pages which contained a reference to a name from the staff list. The passages relating to a person were concatenated to form a surrogate document for the person. P@NOPTIC indexed the surrogates and, in response to a research-topic query, would return a ranked list of experts, including their contact details and links to the evidence supporting their expertise on that topic. Nick and co-authors presented a paper on `P@NOPTIC Expert` at the Ausweb conference in 2001[12]

`P@NOPTIC Expert` made a great demonstration, but was derided by some CMIS researchers. Researchers who spurned the web – "I publish in prestigious journals and don't waste my time putting fluff on the web!" – found that they were ranked lower than they deserved, and sometimes science communicators like Tom McGinness, Carrie Bengston, and Janelle Kennard, would rank ahead of real experts. That was because a lot of the web content within CMIS was actually

[10] Due to Kalervo Järvelin and Jana Kekäläinen.

[11] `https://david-hawking.net/pubs/bailey_ipm03.pdf`

[12] `https://david-hawking.net/pubs/craswell_ausweb01.pdf`

written by communicators and listed them as the people to initially contact.

Returning communicators as contacts for research topics actually makes a good deal of sense for many use cases. For journalists looking for expert comment, and industry partners looking for expertise, contacting a communicator would be an effective if indirect place to start.

4.6 *The TREC-2001 Web Track

In this edition of the Web Track, there were again two tasks: a Topic Relevance Task, and a Homepage Finding Task. Both used the WT10g corpus. The topic relevance task was very similar to past editions of TREC ad hoc. The homepage finding task used 145 queries chosen by NIST assessors. The assessor would find a homepage within WT10g and compose a query to retrieve it. On this task, the TNO/UTwente team of Wessel Kraaij and Thijs Westerveld made use of prior probabilities derived from URL lengths and showed that a submission using only URL length priors performed almost as well as one which used only anchor text. The best performing content-only system achieved only about 30% of the score of the "winning" system.

4.7 *SIGIR Web Search Tutorials and 9/11

The annual SIGIR conference starts with a tutorial day, during which attendees can, for a fee, learn about specialised information retrieval topics. In early 2001, I was approached by Andrei Broder (Chief Scientist at Alta Vista) to join him and two other very distinguished people, Prabhakar Raghavan (Chief Scientist at Verity, a P@NOPTIC competitor) and Krishna Bharat (a Research Scientist at Google), in offering a full-day tutorial on Web Search. My contribution was to be on Evaluation.

2002 part of the audience for our Web Search Tutorial in Tampere. Prabhaka Raghavan (left) and Krishna Bharat sit in the front row, while Andrei Broder lectures and I take photos.

SIGIR 2001 was held in New Orleans and it had a very different flavour to other SIGIRs. Jazz bands played in the N'Awlins airport. In places, sewage bubbled out onto the side walks – the streets were lower than the water in the nearby Mississippi River and Lake Pontchartrain. We ignored warnings from hotel staff that we were not safe to walk down Canal Street to the conference venue. They told a group of ten conference-goers that they would be risking their lives if they went to a jazz concert across Canal St. On the other hand, Keith van Rijsbergen toured the city on public buses and was unfazed by being the only white face on board – white people didn't use public transport.

The antique shops and classic buildings in Royal Street in the French Quarter were a haven of calm civilisation, but one street over Bourbon Street fizzed with bars and sex joints. "Two ladies for every gentleman! I know how you do like titties!"

Nick Craswell and I made sure we ticked off the tourist must-sees (not in Bourbon Street!) We drank Hurricanes at Pat O'Briens and queued up to listen to spontaneous jazz music at Preservation Hall – a run-down house with the internal walls knocked out. The musicians played continuously and the audience flowed through, dropping a small donation to spend as long as they liked in the packed space. We ate catfish and jambalaya. I went to a voodoo house and visited the cemetery where bodies are sunbaked in the walls before burial in the waterlogged ground.

Our Monday tutorial was very well attended. People were very excited by web search engines and wanted to know how they worked.

Unfortunately, the conference opening the next day was mainly attended by people like me, those who didn't watch American television and didn't know about the 9/11 atrocities in New York and Washington DC. US attendees, like the rest of the country, were in deep shock. A wild-eyed man grabbed the microphone and exhorted us to head to the nearby Tulane Medical Center to donate blood. George W Bush was nearby and his jets disappeared in a hurry. Otherwise there were no aircraft in the sky.

Lack of aircraft posed a significant problem for attendees getting home. Some Americans rented cars and drove thousands of kilometres. We Australians with tickets on United Airlines / Ansett were well and truly stuck. United Airlines were in shock at the loss of their staff, aircraft, and passengers, and seemed unlikely to resume domestic flights "any time soon." Flights over the Pacific seemed weeks away. Ross Wilkinson took responsibility for getting his team members, Nick Craswell and me, home. We decided to try to first solve the problem of getting to Los Angeles. Ross found one-way tickets to LAX on Delta for USD2000 each! He decided to take them but the connection dropped out before he could pay. He then successfully tried South-West Airlines who I believe may even have accepted our United tickets. Not only that but he also found seats on an Air New Zealand flight to Sydney, with a good connection in LAX. Thanks Ross!

We had four unplanned days in N'awlins. We spent one of them driving down to the Gulf Coast with Justin Zobel and Alistair Moffat, admiring the execrable puns that people used when naming their beach houses, and contributing even worse ones ourselves.[13]

N'Awlins airport was eerily deserted, with no sign of a jazz band. Security was intense. I had to stand on a pad with my trousers falling down, while being patted down for weapons. Everything in my suitcase was taken out and inspected before being much more neatly repacked. They confiscated a tiny six-inch steel ruler which my father-in-law had acquired in the Army Engineers in 1941. I was very unhappy about that but I understood the situation.

Amazingly, LAX lived up to its name, with no additional security – my memory is that we just walked onto the plane. Arriving there, Ross bowled up to the Air New Zealand desk, and said,

– "I have a million Ansett frequent flyer points, please upgrade us all."
– "I'm sorry sir, you'll have to contact Ansett for that."

As they were well aware it would be impossible to do that, since Ansett (owned by Air New Zealand) had just declared bankruptcy. Anthea Roberts, a Canberra manager with Ansett and later a CSIRO HR manager, was one of the victims of the collapse. She was still upset about that, years later.

Two other major Australian companies went bankrupt within a couple of days, leading to the probably apocryphal hard-luck story of the person trying to phone HIH insurance on their OneTel phone to make a claim on their travel policy after the collapse of Ansett! We were delighted that Qantas accepted our Ansett tickets for the domestic sectors beyond Sydney.

Tampere, Finland: Andrei, Prabhakar, Krishna and I reprised our Web Search tutorial at SIGIR 2002 in Tampere, Finland. It was again quite successful, though I had more fun being a disruptive presence at the XML retrieval workshop later in the week.

[13] Alright, I exonerate Ross, but I swear I wasn't the only one.

4.8 *The TREC-2002 Web Track

By 2002 the crawl on which WT100g and its derivatives were based was five years old. During that time web publishing had significantly changed. Thanks mostly to style sheets and JavaScript, the average size of webpages had roughly tripled. Accordingly, this edition of the Web Track used a 1.25 million page, 18 gigbyte crawl of the `.gov` internet domain. The crawl was carried out by Ed Fox's team at Virginia Tech using software supplied by Charlie Clarke of the University of Waterloo. Ian Soboroff of NIST shipped CSIRO a hard disk of the raw crawl which, due to very large files and binary formats from which text content was extracted, was nearly four times as large as the version eventually distributed. The disk was formatted in XFS and a little bit of fiddling by Peter Thew was needed to read it on our machine `croc`. After extracting text from binary formats, excluding very large files and converting to TREC format, we (CSIRO) distributed .GOV and its link connectivity data on eight CD-ROMs.

The two Web Track tasks in 2002 were a Topic Distillation task and a Named Page Finding task. The topic distillation task didn't seem to be properly understood by judges or participants and not much was learned. In named page finding, the targets were not home pages, but rather specific documents, such as *2001 Carbon Emissions By State*. The query was essentially the title of the page. Unsurprisingly URL length was not a useful prior on this task, but anchor text and structural information seemed to help.

4.9 *The TREC-2003 Web Track

The .GOV corpus was used again, and an interactive task organised by my colleagues Ross Wilkinson and Mingfang Wu was added. The non-interactive tasks included a Topic Distillation task and a mixed Named page / Homepage finding task. The topic distillation task was considered more realistic than the previous year's. In the other task, some groups tried to build classifiers to determine whether a particular query looked for a home page or for a named page. Others trained ranking functions to work on the mixture of query types.

2008: My CSIRO colleague Mingfang Wu, in Tasmania for ADCS 2008

For the purposes of the Interactive task, NIST hosted a P@NOPTIC installation configured to provide two search interfaces, one using a ranking function which used web features such as anchor text, and the other did not. P@NOPTIC copped heavy flack from senior people like Bill Hersh and Nick Belkin, because retrieving the home page of one key government site (`http://irs.gov`?) at

rank one required a URL redirection which we somehow missed. Fortunately, by that stage, I had developed a reasonably thick skin.

4.10 *The TREC-2004 Web Track

This was a slight extension of the 2003 Web Track. A mixture of 75 queries of each type: topic distillation, home page finding, and named page finding were to be run over .GOV. Apart from the main task of achieving high accuracy retrieval, another task was to build a classifier to group the queries into their types.

2004: At SIGIR on the campus of Sheffield University. Ian Soboroff (NIST) and Nick Craswell (Microsoft) communicate intently. Thijs Westerveld (red and green shirt) takes a more relaxed attitude.

2005: SIGIR was held in Salvador, Brazil. From left: Bruce Croft (UMass), Kathy Griffiths (ANU), Manmatha (UMass), and me. ***Photo: the waiter.***

By 2004, I felt that the TREC Web Track had achieved considerable success, but had gone about as far as it could go within the constraints of what was feasible. It had shed light on: differences between

web search and TREC ad hoc retrieval; different types of web search; and the ranking features useful in satisfying different types of information need. To fully succeed in web search it seemed that you needed a complex ranking function, and masses of training data to set its parameters. There was no practical way that academic researchers working in TREC could obtain the large scale user interaction data needed for this. Bruce Croft's group at UMass set up a search engine front-end to collect user data but the amount of data collected was small.

In 2006, in a well-intentioned move to support academic research, AOL released a query log.[14] Users were anonymised but retention of user-ids allowed all the queries submitted by an individual to be identified. A reporter from the New York Times managed to identify Thelma Arnold, from her sequence of queries. The NYT article caused a scandal, triggering lawsuits, causing the departure of three AOL employees, and virtually eliminating the possibility that any search company would be willing to release user data to the general research community. One of the departees was Abdur Chowdhury, who I knew reasonably well. I felt very sorry for him because the query log was scientifically useful, and its release was done with the best of intentions. If contractually limited as the TREC data sets were, no harm would have been done.

Abdur Chowdhury's career doesn't seem to have suffered too badly from this setback. He was Chief Scientist at Twitter and now holds a CEO position.

I didn't attend TREC in 2004, and didn't return until the 25th anniversary in 2016. I owed my search career in large part to TREC, but it was time to look beyond it.

In later years TREC continued to work with very large web corpora (in other tracks), notably GOV2 (a far larger crawl of government data, distributed by CSIRO), ClueWeb09 and ClueWeb12 (crawled and distributed by Jamie Callan's team at CMU). More recently, the TREC Web Track has been revived with Nick Craswell as a co-ordinator and with promise of new learnings.

I had decided to concentrate on Enterprise Search, which, despite its much smaller scale, was in many ways at least as diverse and challenging as mainstream web search. I had some hope that it might be possible to research some aspects of enterprise search within the context of TREC, but realised that, through P@NOPTIC, I had access to user data and other resources which could never be provided within TREC.

4.11 *2005 and beyond: The TREC Enterprise Track

At SIGIR 2004 in Sheffield, I started discussions with the TREC organisers, Donna Harman, Ian Soboroff, and Ellen Voorhees, about how a suitably realistic enterprise search test collection might be set up.

Over the next months, we considered the possibility of using the enterprise data from failed companies. Some researchers had already made use of the Enron data released by a US court. I was offered data from a number of bankrupt companies, presumably for a price, but there were two problems with proceeding down this path:

1. There seemed to be no feasible way to generate representative queries or information needs relating to the documents from a failed company.
2. Ethical concerns. The people who created the documents (possibly including personal emails) within the company had not given permission for those documents to be viewed outside the company.

In the end, the 2005 Enterprise Track made use of a public mailing list published by the World Wide Web Consortium (W3C). This didn't seem very relevant to the problems being faced by P@NOPTIC.

The Enterprise Track continued and Peter Bailey created a better test collection for use in 2007. It was the CSIRO Enterprise Research Collection (CERC). It was based on a crawl of `csiro.au` and a set of realistic information needs devised by CSIRO's Science Communicators based on public requests for information. Creating this collection was a significant achievement on Peter's part, but although the retrieval scenario was realistic, it represented a very narrow use case.

[14] `https://en.wikipedia.org/wiki/AOL_search_data_leak`

4.12 ∗CSIRO Search Research Pursued Outside TREC

This section is organised by research topic but the topics do appear in roughly temporal order of when the research commenced. Many of the projects were carried out in collaboration with ANU PhD students.

4.12.1 ∗Distributed Information Retrieval

Distributed information retrieval covers the case where a retrieval index is partitioned across a number of servers and a search broker is used to deliver a single set of results in response to a query. In the simplest case, the servers co-operate by sharing information with each other and/or with the broker.

In the more challenging uncooperative case, five research problems may arise:

1. How to discover the existence of search servers;
2. How to characterise the holdings of the available search servers;
3. How to select a suitable subset of available servers;
4. How to translate the broker's query language into the language of each of the servers; and
5. How to merge results from multiple servers into a high-quality ranked list.

My own PhD was titled *Text retrieval over distributed collections*. It covered both the cooperative (AP1000 parallel retrieval, and retrieval from a COW (Cluster of Workstations) and the uncooperative cases. In the latter, I modeled uncooperative servers using the nodes of the AP1000, and developed the method of lightweight probe queries to select the most useful servers to answer the query. PADRE's Z-mode scoring (see Page 27) made the merging of partial results lists a trivial problem because Z-mode scores are independent of collection statistics.

Nick Craswell's thesis *Methods for distributed information retrieval* presented and systematically evaluated a wide set of result merging and server selection methods. Nick and I supervised Jared Cope's fourth-year honours project on search server discovery. Jared developed a search interface classifier and ran it over a large crawl to identify all the independent search interfaces. The work was published in the 2003 Australian Databases Conference[15] and was written up in the ANU Reporter. It achieved substantial international interest, and a Japanese organisation funded Nick to travel to Japan to explain, discuss and advise on the work.

In 2003 I collaborated with Yves Rasolofo and Jacques Savoy of the University of Neuchâtel to build and evaluate a *current news metasearcher*.[16] Later, I travelled to Neuchâtel to examine Yves's thesis – in French.

Another of my ANU PhD students, Paul Thomas, also wrote a thesis in Distributed IR: *Server characterisation and selection for personal metasearch*. He and I initially simulated a more realistic distributed IR scenario for the web. We divided a web test collection (WT10g) into a very large number of servers – the actual servers which had been crawled. If I remember correctly, we assumed that all the tiny ones were crawled and aggregated into a single search service, and that each of the others operated its own independent search service. We proposed algorithms based on anchor text to determine which of the servers should be included in the metasearch for each incoming query. Those algorithms worked very well.[17]

At various stages over the three decades of the Web, distributed IR has seemed to be a promising solution to problems such as the perceived inability of search engines to cover the entire web, or the inability of search engines to cover the dark web. Each time it has turned out that a centralised search engine has been able to effectively satisfy a very high proportion of user needs for web search, and the case for uncooperative metasearch has faded away. Mind you, centralised search engines like

[15] https://david-hawking.net/pubs/cope_adc03.pdf
[16] https://david-hawking.net/pubs/rasolofo_ipm03.pdf
[17] https://david-hawking.net/pubs/hawking_sigir05.pdf

Google, Bing and Baidu are in fact built on cooperating distributed search servers. A single query may in fact be processed independently on more than a thousand servers each responsible for a small partition (shard) of the total index. A broker is responsible for distributing the query and merging the results.

Another way in which web search engines rely on distributed IR techniques is in composing the result page (SERP[18]) from results obtained from different types of search index: web, images, news, maps, videos etc. In composing a SERP decisions must be made about which indexes should be used, which type of results should be included and where the results of a particular type should be included in the SERP.

The main focus of Paul Thomas's thesis was a potentially very useful application of Distributed IR techniques, not amenable to a centralised solution. Paul built a search broker called `PIS` (Personal Information Service) designed to be used by a knowledge worker. A typical knowledge worker may have access to a large number of different search services which can't feasibly be combined into a single monolithic index. They may have a search over their own personal documents, various local search services over corporate data, search services on cloud services like SalesForce, general web search services such as Google or Bing, and external proprietary search services. PIS was an attempt to provide a single search broker to access all these different services, intelligently deciding which services to consult for each input query.

`PIS` needed to run on both Windows and Linux platforms so Paul learned C# in order to facilitate portability. He used his `PIS` apparatus to compare server selection algorithms in this environment, using side-by-side evalauation.

Paul has gone on to a distinguished research career,[19] first at CSIRO and then at Microsoft Bing. Some of his work is described in the following sections.

4.12.2 *New Evaluation Methodologies

CSIRO's requirements for search quality evaluation were not met by the TREC ad hoc paradigm. Over the years, we developed and used a number of alternative measurement and comparison methodologies.

*Side-by-side Evaluation

As recounted on Page 116, P@NOPTIC had been using side-by-side comparison to demonstrate the merits of P@NOPTIC over an incumbent search engine. We soon realised that, by adding voting buttons or scoring sliders, a side-by-side tool could be used to scientifically measure the relative performance of two engines or different tunings of the same engine. To do valid comparisons, the tool needed to randomise the left-right placement of the two panels. Paul Thomas's version did this, and added `I prefer this one` buttons on each panel, and `equally good` and `equally bad` buttons between them. A simple *sign test*[20] can be used to determine whether the results of many side-by-side trials indicate a statistically significant preference for one engine over the other.

Paul and I did a validation experiment for this tool,[21] persuading users to conduct their normal web search activities via his search interface rather than by going directly to a search engine. This is an ideal way to do an evaluation, because the information needs are real, and because judgments are made at the time of the search, and in full understanding of the need behind the query. We compared real Google results lists with various modifications of those lists. Our conclusion was that the side-by-side method achieved useful and reliable results. The method wasn't sensitive to very small changes but revealed differences which mattered to users. Unlike traditional offline test collection evaluations, comparisons could, if desired, take into account all the features of the SERP,

[18] Search Engine Result Page
[19] `https://scholar.google.com/citations?user=B7c-WcAAAAAJ`
[20] `https://en.wikipedia.org/wiki/Sign_test`
[21] `https://david-hawking.net/pubs/cikmfp633-thomas.pdf`

not just the document ranking – e.g. snippets, spelling suggestions, related queries, advertisements, carousels, etc.

CSIRO used variations of side-by-side evaluation in other studies such as Alex Krumpholz's study of medical literature retrieval, described on Page 73.

*Queries from Site Maps and Lists

I'm not sure how common it is these days, but in the first decade or two of the world wide web, many web sites included a site map, providing links to the best resource for each of the topics identified as important. These site maps often provided a source of (query, best-answer) pairs[22] which could be used for low-cost evaluation and tuning of a search services. We assumed that the entries in the site map were the topics considered to be important by the web site publisher.

The first time we used this type of evaluation we used a members list rather than a site map. In order to evaluate the ability of web search engines and directories to process navigational queries, we used the list of members of the International Air Transport Association (IATA). The name of the airline obtained from the IATA list was the query and the home page of that airline was the required answer. Results are presented in the graph on Page 54.

Page 67 includes a description of the site map method used by Trystan Upstill to compare internal and external search of Stanford University.

We used query sets derived from site maps quite often in the early days, both for research and for quality control of P@NOPTIC demos.

*C-TEST: CSIRO Toolkit for the Evaluation of Search Technology

In the TREC ad hoc retrieval paradigm, the quest is to find relevant documents, not necessarily useful ones, and full credit is given for retrieving documents which convey the same meaning as others already retrieved. This paradigm may make sense in information seeking to support academic research or intelligence gathering, but it doesn't model people searching for the documents which will allow them to carry out tasks such as building a backyard water feature, enrolling at a university, or registering a dog.

Another TREC paradigm is known-item search. Typically the task is to re-find a document that you have seen before. Alternatively, you know, or suspect, that there is a single page which will give you the information you want or which will enable you to navigate to the information you want. TREC tasks along these lines included a low-cost way of assessing the performance of retrieval over OCR-ed documents, homepage finding, named-item finding.

At CSIRO, we developed a paradigm, represented by an XML testfile format and a set of tools, to cover situations where:

1. *Usefulness* rather than *relevance* is what matters;
2. Queries may have more than one interpretation. E.g. `Java` may represent information needs relating to an island, a programming language, or an American term for coffee. Interpretations are associated with a weight, such as relative prevalence.
3. Each interpretation lists a handful of topic distillation answers meeting the need behind the interpretation.
4. Each answer has a weight reflecting its value to satisfying the need behind the interpretation.
5. Queries may be weighted relative to each other, e.g. to reflect their prevalence in the query population or their value.
6. Every effort is made to avoid giving credit for duplicate documents. Multiple URLs can be marked as *equivalent*. Credit is only awarded for one answer in the equivalence set.

The format and tools were published[23] and made available to others. Tom Rowlands used C-

[22] The link anchor is the query and the target of the link is the best answer.

[23] `https://david-hawking.net/pubs/hawking-rowlands-thomas09.pdf`

TEST in his PhD work under my supervision, and Funnelback used it in tools for tuning a search installation.

Thinking about conditional utility

Early in his PhD, my student Tim Jones organised an experiment in which ANU students attending Orientation Week were lured into a computer lab and persuaded to undertake search tasks, while being monitored. One of the tasks was:

> Find three micro-breweries operating in Canberra.

Simplifying things slightly for clarity, available search results included pages A, B, and C for individual breweries, a page ABC which listed all three, and a page AB which listed two of them. You could say that, judged in a vacuum, A, B, and C might score 1, ABC 3, and AB 2. However, if ABC is returned at the top of the search results, the task is complete and the *conditional utility* of all following pages is close to zero.
On the other hand, if the result ranking goes B, C, AB, ABC, ... the conditional utilities go 1, 1, 1, 0, ...
I certainly thought about conditional utility when working on C-TEST but I can't remember whether I came up with a solution.

*Where's the population?

My wife Kathy Griffiths accompanied me to SIGIR 2005 and attended several of the talks. As a researcher in psychology and mental health, she relied heavily on statistical analyses. During Mark Sanderson's talk on statistical analysis of information retrieval systems, she kept demanding, "Where's the population?" Her point was that, because the topics in TREC ad hoc and other evaluations were artificially created by NIST assessors, they were not a representative sample of any real population of topics.[24] System A may outperform System B on TREC ad hoc but statistical inference cannot be used to make any interesting general conclusion about performance. We cannot be confident that System A will work better than System B on TREC topics in general, or in other search settings, such as web search or search within an organisation.

I realised that, in P@NOPTIC installations, we had clearly defined and interesting query populations – the search workload of each installation. If we defined the population as the entire set of queries submitted to a P@NOPTIC service within a period of time, and tuned parameters on a random sample of that population, then we would have a statistically valid expectation that performance on the sample would reliably predict performance on that workload. Making the reasonable assumption that search workload changes only slowly over time, tuning P@NOPTIC on last month's workload, should achieve good results today.

Tom Rowlands, Ramesh Sankaranarayana and I presented a poster at SIGIR 2007 in Amsterdam on *Workload sampling for enterprise search evaluation*,[25] in which we showed substantial deviations from results obtained by sampled queries when query sets were obtained by other methods, such as site maps.

Ethics approvals for human experimentation

We conducted quite a few studies involving human volunteers. All of them involved an ANU PhD student and in each case the experimental protocol was submitted for approval by the ANU Ethics committee, prior to commencing the study.

[24] Although multiple assessors (not sure how many) were used, it is doubtful that they were a representative sample of the population of intelligence assessors. In any case, the topics they constructed were very unlikely to be a representative sample of topics researched within US intelligence agencies. Further departing from random sampling, topics estimated to retrieve a number of relevant documents outside a target range were discarded.

[25] `https://david-hawking.net/pubs/rowlandsHS07.pdf`

***Post my departure from CSIRO**

Paul Thomas finished his PhD in 2008 and headed off for R&R in Senegal and The Gambia. Unfortunately, a medical condition saw him medevaced back to Australia, where he started a postdoc at CSIRO. After completion of the postdoc he became a research scientist in CSIRO and remained there until 2016, when he joined me and Peter Bailey in Nick Craswell's science team at Bing.

2008: Paul Thomas celebrating completion of his PhD at the old Wig and Pen.

In his time at CSIRO, Paul recalls working on:

- Applications of metasearch.
- Evaluation, e.g. TREC enterprise and contextual suggestion tracks; also increasing amounts of thought on how to do evaluation.
- How people navigate websites to find information.
- Effects of screen size/form factor on searcher behaviour.

With Tom Gedeon, Paul supervised an ANU PhD student Jaewon Kim. They had access to eye-tracking equipment and used it in studying the last two items on that list.

4.12.3 *Value of Link Evidence in Enterprise Web Search

Trystan Upstill's thesis *Document ranking using web evidence* explored to what extent web evidence such as inlink counts, PageRank, and anchor text can improve the quality of search in small webs. An interesting conclusion was that while small benefit was achievable from query-independent evidence, there seemed to be no benefit to using the far more sophisticated PageRank method over simple inlink counts.

Trystan, Nick Craswell and I were also very interested in evaluating the role of external evidence in responding to search requests scoped to an organisation or to an individual web site.[26] Is an

[26] `https://david-hawking.net/pubs/hawking_adc04.pdf`

external search engine like Google able to provide better scoped search of an organisation because of its ability to use link evidence from the web as a whole? Stanford University (`stanford.edu` served as a case study because, unsurprisingly, Stanford was one of the first adopters of the Google Search Appliance (GSA). We compared more than 1200 navigational queries derived from Stanford's A-Z site directory. We did a comparison of the search performance of Stanford's GSA with that of `google.com` when adding a `site:stanford.edu` restriction.

Trystan did most of the work on this experiment. He eliminated some queries for which the answer was not reachable by both search services. Examples included internal-only content, and answers outside the `stanford.edu` domain. The next issue was to expand the answer sets to include all the equivalent URLs for each right answer. Sometimes simple URLs such as `department.stanford.edu` might redirect to deeper URLs such as `department.stanford.edu/content.cgi?page=1`. At the time `google.com` tended to return the shorter version, while the GSA tended to return the ultimate redirection.

Having done all this, the result sets for both engines could be fairly evaluated. In the very common case where both engines retrieved a right answer at rank one, no manual judging was needed. In other cases, Trystan manually looked at the two result sets to make sure that an unrecognised right answer wasn't present above the answers the evaluation script found in the results lists. After taking all this care, there was no significant difference between the two engines. In 907 out of 1266 queries, performance was identical; in 194 cases the appliance performed better; in 165 cases the difference was in favour of the whole-of-web service.

In the same paper I compared the internal and external anchor text for a number of different Australian organisations. Francis Crimmins had performed a large crawl of the `.au` domain and I extracted all the anchor text referencing pages from the sites I studied. I found that there was a far larger quantity and diversity of the internal anchor text. There were a lot of external links to Australia Post, for example, but nearly all of it was targeted at the organisation's homepage or at the postcode directory. Most of the external anchor text consisted of minor variations on `australia post` and `postcodes`. Similar patterns applied to other organisations such as banks and universities.

We found negligible improvement to within-organisation searches when internal anchor text was augmented by external anchors. External anchor text was of no help answering the vast majority of within-organisation queries.

4.12.4 *Near-Duplicate Elimination and Result Set Diversification

TREC ad hoc evaluated the value (relevance) of every document independently. There was no penalty if the facts or opinions in one retrieved relevant document overlapped with those in other retrieved relevant documents. In the worst case, a retrieval system might gain credit for retrieving multiple copies of the same document!

In real-world applications, searchers are not happy when their result list is full of repetition and overlap. A classic case was the pre-P@NOPTIC search at Westpac, where `home loan` retrieved ten apparently identical "archived media releases."

Elimination of duplicates is easily achieved using checksums. It is best done during the document gathering phase to reduce unnecessary cost and delay in gathering and indexing as well as improving results. Unfortunately, web pages often include a timestamp indicating when the page was accessed, a copy of their own URL, an incrementing visitor count, or headers, footers, and stylesheets specific to a segment of the site – identical information being published with different branding. Web search engines and researchers devoted considerable energy to recognising and suppressing near-duplicate pages. Francis Crimmins developed crawler heuristics to achieve this in P@NOPTIC.

Elimination of duplicates is an important start, but there may be complete redundancy in the information conveyed by two documents even if they are written in different words. In 1998, Carbonell and Goldstein introduced the concept of maximal marginal relevance (MMR). When applied to search, MMR is the vector-space distance (see the panel below) between a candidate document and the centroid of the set of documents already retrieved. The retrieval system first creates a ranked list of k results. The document at rank one stays where it is, but documents lower than that are re-ranked

using a combination of their similarity to the query and the MMR distance between them and the documents above them.

The Vector Space model of information retrieval

Many decades ago, Gerard Salton of Cornell University proposed a retrieval model in which documents and queries were represented as high-dimensional vectors. Each dimension in the vectors corresponded to a word in the vocabulary. If a term were present in the document (or query) the corresponding entry would receive a non-zero weight, otherwise the vector entry would be zero. In this vector space model, ranked retrieval was achieved by calculating the distance of each document's vector from the vector representing the query, and sorting by increasing distance.
PADRE used probabilistic and lexical distance models rather than vectors.

The sort of overlap addressed by MMR was not a significant problem in P@NOPTIC projects, and we never implemented anything like MMR. However, when the search covered multiple repositories or web sites, results rankings could be unfortunately dominated by one source. In the initial version of P@NOPTIC search at CSIRO, a search for `Peter Corke` ranked a couple of his home pages first, but then returned more than forty messages from a robotics mail archive that he moderated. Those mail messages pushed other useful content way down.

To solve this and similar problems I implemented a configurable *same site suppression* mechanism in PADRE. Scores for results from "sites" already present in the ranking were attenuated by a factor which increased with the count of results from that "site". A site could be defined in the configuration to be a server, a top level directory or a subsidiary level directory. The number of results from a "site" before attenuation kicked in was also configurable.

4.12.5 *Searchability

There's a flipside to understanding how search engines work – you can use the knowledge to promote web sites in search engine rankings. That's mostly an issue on the open web, where company profits can depend heavily on the ranking of their web content in response to queries relating to their business.

It's an issue within some organisations too. On Page 163 I describe the competition for visibility within the domains of the London School of Economics and Political Science.

A paper at the Australasian Document Computing Symposium in 2002 (ADCS 2002 – *Buying bestsellers online: A case study in Search & Searchability*[27]), led by Trystan Upstill, looked at the visibility of best seller books in search engine results. Our study found that there were dramatic differences in visibility across online bookstores selling the same books. It concluded that the visibility of a page from which a particular book could be purchased depended upon the form of its URL. A complex, obscure, and ephemeral-looking URL reduced the chance of external links (and anchor text) to that page. Furthermore certain patterns of URLs reduced the chance that book pages would be included in search engine indexes.

Another Trystan-led paper at ADCS2003[28] used data reported by Google and FAST to investigate various possible biases in PageRank scores, and to check whether PageRank added value above and beyond simpler inlink counts. At the time, the Google toolbar reported PageRanks as a number in the range 0 – 10, and they also represented it visually in the length of a coloured bar. Trystan found that the bar was more informative than the number, since there were 41 possible lengths of the bar. We concluded that there were biases toward large multinational companies and tech companies, but no particular bias toward US companies. We also concluded that there was no particular value to PageRank over the logarithm of inlink count – a finding made more interesting by Trystan's subsequent employment at Google!

[27] `https://david-hawking.net/pubs/upstill_adcs02.pdf`
[28] `https://david-hawking.net/pubs/upstill_adcs03.pdf`

4.12.6 *Value of Subject Metadata

At ANU in the 1970s, the only way to retrieve information by subject was to use library classifications, such as Dewey Decimal 537 – Electricity and Electronics, or to look up a *subject* card catalogue using indexing terms. If you were a medical researcher you may have had access to the MEDLARS system operated by the US National Library of Medicine (NLM). The NLM carefully, and at great expense, constructed a thesaurus of medical terms, known as MeSH (the Medical Subject Hierarchy) and used them to tag medical research papers. Researchers could send a query to NLM, comprising a set of MeSH terms, and have printed abstracts of matching papers returned to them.[29] In those days, indexing terms were essential because it was not technically feasible to index the full text.[30]

Indexing terms used to label a document were a type of *subject metadata* – data about the document's subject.

Many librarians believed that subject metadata for web pages would be the only effective way of ensuring that web content could be retrieved by subject. At the second World Wide Web conference in 1994, the Dublin Core Initiative began. Dublin Core defined a set of metadata elements such as DC.title, DC.subject, DC.author and DC.date which were recommended for inclusion in all web pages. A senior CSIRO librarian / IT person told us that P@NOPTIC should retrieve content solely on the basis of those Dublin Core elements. She said that the fact that most of CSIRO's web content would consequently not be retrievable, would encourage/force CSIRO publishers to insert the relevant DC tags! ☹

Justin Zobel and I conducted a study (*Does Topic Metadata Help With Web Search?*[31]) on the web pages published by RMIT University, an organisation with a very strong commitment to the use of metadata. We found that, in practice, metadata was very sparse, and of poor quality. It was clear that many of the people applying metadata didn't understand what purpose metadata was supposed to serve, and were very poorly motivated. Metadata strings were very frequently copied from other pages, whether applicable or not, and quite often nonsense was entered.

Looking at the queries submitted to the RMIT search engine, we found that only a negligible proportion of queries could be effectively answered using metadata, and in almost all of those cases, the content of the document provided results which were at least as good, often much better.

I presented our study at a meeting of metadata professionals, and it went down like a lead balloon. ☺

4.12.7 *Quality-Oriented Information Retrieval in a Health Domain

My wife Kathy Griffiths was a researcher at ANU's Centre for Mental Health Research (CMHR). She and colleague Helen Christensen had evaluated the quality of mental health information published on the web, and found that much of it was poor quality, and some even harmful. They had developed a 20-point evidence-based rating scale for the quality of information on web sites covering the topic of depression. A very high quality source of information about depression should cover multiple treatments, recommend treatments for which there is high quality evidence based on randomised trials, and not recommend treatments lacking such supporting evidence.

Many web search experts regarded PageRank, or other measures derived from link recommendations, as measures of the quality of a web page. CMHR wondered whether PageRanks would accurately predict evidence-based ratings.

CSIRO and ANU Computer Science collaborated with CMHR on a number of projects in this area. Thanh Tin (Tim) Tang was a PhD student supervised by me, Nick Craswell and Ramesh Sankaranarayana. We were not at all sure that the links on which PageRank relies would necessarily flow predominantly to evidence based sites.

[29] Before that, medical researchers consulted hard copy volumes with numbered papers listed under index terms.

[30] Ev Brenner told an Infonortics Search Engines Meeting that despite countless efforts in government and industry to build thesauri, MeSH was the only one which had succeeded, and creating it had cost the US taxpayer huge amounts of money.

[31] `https://david-hawking.net/pubs/hawking_zobel_jasist.pdf`

We ran an experiment to test this, and also tried to devise our own method for automatic quality assessment. We crawled a large number of candidate sites and used expert human judges from CMHR to evaluate their quality, using a scoring sheet which resulted in quality scores in the range 0 (hopeless) to 20 (perfect).

2006: Tim Tang's PhD graduation. From left: Me, Tim, and Ramesh Sankaranarayana. ***Photographer unknown.***

Using a subset of the data, Tim Tang used a relevance feedback method to learn a complex, weighted query. Running that query against an index of all the pages of all of the sites, and aggregating the page scores for a site, then scaling to the desired range, resulted in site scores which could be compared to the human quality scores. We found that scores for this Automated Quality Assessment (AQA) method correlated remarkably highly with the human assigned scores ($r = 0.85$). Correlation between PageRank and the human assigned scores was much weaker. The work was written up as *Automated assessment of the quality of depression web sites*.[32]

In another study we ran a set of 51 depression-oriented queries and 50 queries consisting of the names of possible depression treatments, against various search engines, and judged whether each result recommended for or against the treatment. All queries were judged for relevance on a multi-point scale, while the results for treatment queries were judged as to whether they recommended the treatment, recommended against it, or made no recommendation. A key feature of the study was that we could use non-expert judges. The study was published as *Quality and relevance of domain-specific search: a case study in mental health*.[33]

*Quality Focused Crawling

Kathy Griffiths and colleagues created an online *What works and what doesn't* site for depression treatments: `bluepages.anu.edu.au`. It used P@NOPTIC to provide the ability to search information on that site and, optionally, to search a collection of manually chosen external web sites providing quality information on depression. The external sites collection was defined by a very long list of seed URLs and hundreds of crawler include and exclude patterns. A lot of work was needed to create and maintain the seeds and the patterns. Could the process be automated?

Chakrabarti et al proposed the term *focused crawler* in 1997 to describe a crawler which prioritised content related to a topic. Tim Tang made three variants of a crawler: one a straight-forward

[32] `https://david-hawking.net/pubs/griffiths_jmir.pdf`
[33] `https://david-hawking.net/pubs/tang_domainspec.pdf`

breadth-first crawl, one which prioritised content relating to depression and the other which prioritised depression content of high quality. Promising results were achieved with the third one and the work was published at CIKM (Conference on Information and Knowledge Management) in 2005.[34]

4.12.8 *Nullifying Spam

In the late 1990s and early 2000s web spam became an increasing challenge to web search engines trying to deliver useful results to searchers. "Black hat" people in the search engine optimisation community gathered knowledge about how search engines operated and used it to manipulate search engine results for their own benefit or for that of their clients. Their activities usually had the effect of reducing the extent to which search results satisfied the need behind the query. Many techniques were used to achieve these harmful objectives. To illustrate them, let's consider BH, the hypothetical operator of a pornographic web site PW.

BH promotes the ranking of PW for sex-related queries, by making sure that sex-related words appear frequently in the content, title and metadata. He further promotes PW by persuading other sites to link to PW using sex-related words in the anchor text. All this is fine – it's "white hat" optimisation. But then BH puts on a black hat and contracts with a spam company to set up a *link farm* for him. This is an array of computers representing hundreds of IP addresses on different networks and hundreds of domain names. The company sets up a nest of links within the link farm, linking to each other and to PW, with the goal of boosting inlink counts and Page Rank. So far, not nice but at least it only affects the ranking for sex-related queries.

The next step is to try to present PW in response to queries which are not related to sex. BH knows that `george w bush` is a very popular query (in 2000) and would like to insert PW in rankings for that query. Some people searching for news about politics may well welcome some pornographic distraction and click on PW if it appears in the search results. But how to make this happen?

There are four basic methods:

1. Insert `george w bush` into the content of PW, hiding it by rendering it in the same colour as the background, or displaying it in zero-point font;
2. *Google bombing*. Use the link farm to create masses of `george w bush` anchor text targetting PW.
3. *Cloaking*. Create an authoritative page about George W Bush, with incoming links and anchor text, and a version of PW (PW2) at the same URL. When that URL is fetched by a crawler, show the George W Bush page, when accessed by anyone else, show PW2.
4. *Black hat hacking*. Exploit security bugs in content management or web publishing systems to modify or replace web pages on sites about George W Bush with a version of PW, or with images and links to PW.

The advent of payments for advertisements displayed within web pages resulted in a new form of spam. People would publish useful web pages full of advertisements and then use SEO techniques to promote them. It's still reasonably common to find sites which republish useful information such as Wikipedia pages, transport timetables, Linux manual pages, etc. in the midst of a forest of ads. If the SEO is successful, these pages may rank higher than the primary source of the information, despite being less authoritative and more painful to read. In other cases, a site appears to be a travel agent offering cheap flights to any destination you name but just refers you to real travel agents after exposing you to advertisements and possibly charging the final site for referring traffic to them.

Spam in search engine result pages (SERPs) is harmful to a search engine's reputation and search companies invest large-scale resources into eliminating it. Andrei Broder described the problems of spam elimination and black hat SEO as *adversarial information retrieval*.

Spam elimination was also being studied in the academic world. There were large amounts of spam in TREC web collections. Gord Cormack and colleagues from the University of Waterloo, de-

[34] `https://david-hawking.net/pubs/tang_cikm05.pdf`

veloped a classifier, ran it over ClueWeb09, and showed that the presence of spam adversely affected the results of participants in the TREC 2009 Web Ad Hoc task.

Outside TREC was the Web Spam Challenge, funded by the European Union, and organised by Carlos Castillo. It used crawls of the `.uk` domain (UK2005 and UK2006) and participants tried to classify pages as spam or not spam.

Tim Jones was an ANU PhD student under supervision of me, Ramesh Sankaranarayana, and (at a distance) Nick Craswell. He addressed the adversarial information retrieval problem from a different point of view, challenging the idea that the presence of spam is inevitably bad. If there is spam in a search engine's index but it doesn't appear in search results the adverse effects are only on the cost of running the searh engine and perhaps on the latency of responding. His thesis was about nullifying spam (i.e. avoiding its harmful effects) rather than eliminating it.

Tim Jones at a Funnelback lunch in Dickson.

Tim ran various user studies to try to determine the cost of spam appearing in search results. He developed a toolbar for the Firefox web browser which allowed arbitrary manipulation of SERPs. For example, you could remove two search results out of ten from Google and insert spam or irrelevant content at arbitrary places in the ranking. Some of the experiments required side-by-side comparisons of manipulated and unmanipulated rankings. Some of them looked at the time taken to complete a task with manipulated versus unmanipulated rankings. Some were conducted in situ, where participants used the Jones toolbar instead of their normal search engine, and their actions were logged. Unfortunately Tim, like other researchers using experimental interfaces, found that participants typically used the experimental system for a couple of days and then drifted back to their normal search engine.

The toolbar is described in `https://david-hawking.net/pubs/jones10.pdf`.

Tim found that searchers care more about how useful a page is, than whether it is spam. A spam page containing useful relevant information is valued more highly than non-spam pages which provide no useful information.

4.12.9 *Medical Literature Retrieval

CSIRO were keen to increase the proportion of its research staff who had a PhD qualification. The CSIRO ICT Centre allowed Tom Rowlands and Alex Krumpholz to enrol as ANU PhD students under my supervision, subject to the proviso that the PhD topic aligned with the project(s) they were assigned to as CSIRO employees. That need for alignment became more and more difficult as time went on, due to the annual reorganisation of projects and the staff assigned to them. It seemed like a great privilege to be able to work on a PhD while being paid as a full time research scientist, but the issue of alignment caused stress. In both cases, the PhDs took much longer than the others I had supervised, largely because of project distractions.

Alex Krumpholz initially started working on the problem of how to exploit document structure (particularly XML structure) in improving retrieval effectiveness. His assignment to an e-health group within CSIRO caused his PhD topic to evolve into *Structural aspects of medical literature retrieval*.

Alex is a person with a well-developed sense of humour and a liking for practical jokes. He was well known in CSIRO for jumping up after rain and shaking a branch in order to drench his colleagues. He told me that while living in Austria, he used to go scuba diving in Austrian lakes. One day, he and his diving mates borrowed a bath tub from someone's front yard, carried it to a lake on the roof of a car, and submerged it upside down, at a depth of ten metres. Using an air hose they filled it with air and dived down to light a burner and cook a steak underwater!

In 2006, Alex and I performed a study of methods for automatically generating a bibliography for a particular topic, given the INEX[35] corpus of scientific articles in XML format. We realised that the presence of citations within the articles allowed the use of anchor text in retrieval. Two anchor text conditions were compared against a condition which ignored anchor text. One of the anchor text conditions used the sentence surrounding the reference as the anchor and the other the surrounding paragraph. Alex built a three-way side-by-side-by-side interface showing the results of the three conditions and asked experts to use a query from their area of expertise and to rate each of the bibliography lists on a ten-point scale. This experiment showed that the use of some form of anchor brought considerable benefit but was unable to determine whether sentence context or paragraph context was better.

We decided to work on this *INEXbib* project very close to the ADCS 2006 deadline. Alex roared into action, writing code, recruiting subjects and consulting with statisticians. After an amazing burst of Alex activity we submitted on time, and we shared the bottle of wine offered as a best paper award.

2008: Nina Studeny, then a CSIRO intern from the Technical University of Vienna, in the old Wig and Pen.

Although Alex often shut himself away in his CSIRO office, surrounded by indoor plants and the latest in Apple technology, it was clear that he found far more enjoyment in working as part of a closely-interacting team, than from working by himself. In 2008, Alex had an idea for how to exploit the structure of XML documents to improve retrieval quality. Nina Studeny was a CSIRO intern for a few months that year, and she and Alex worked like Trojans, analysing document collections, extracting statistics, and building their method. Unfortunately, when the method was evaluated, it showed no benefit. ☹

When Nina was returning to Austria at the end of her internship, I told her that I would be soon attending a meeting of the Information Retrieval Facility in Vienna.[36] She suggested meeting for a coffee and we eventually met up at the stunning Café Central. Her mother had suggested it as a suitable venue for an older foreign visitor. It was a great suggestion, offering a huge array of delicious cakes and coffees.

[35] INEX is a TREC-like initiative, based in Europe and oriented toward retrieval of structured documents, particularly in XML format.

[36] See Section 4.13 on Page 75.

2008: The amazing Café Central in Vienna.

4.12.10 ∗Natural Language Processing

It seemed intuitive that understanding natural language ought to lead to better information retrieval. However, several attempts to demonstrate gains on TREC ad hoc by exploiting NLP (natural language processing) techniques were unsuccessful.

My interest in exploring the value of NLP techniques in information retrieval led me to support an application by Tara McIntosh for a CSIRO PhD scholarship, with me as an adviser. Tara was a student at the University of Sydney and her supervisor was James Curran, a Python guru, whose programming ability I came to respect after encountering him at the Computer Science summer schools which he organised and at which I was occasionally a guest lecturer.

The plan was that Tara would first work on NLP projects and then we would try to apply her techniques to improving our TREC submissions. Tara initially worked on a paper which essentially listed a couple of hundred important established facts in the field of molecular biology, and linked to the papers which established each fact. Tara obtained and downloaded each of the papers, and analysed where and how its fact was stated. The first goal was to determine whether the fact was always stated in the document abstract or conclusion, and thus whether NLP techniques could be applied only to part of the documents, not to the full text. That turned out not to be the case: critical facts were stated in all sorts of odd places including figure captions.

The second goal was to determine what NLP techniques (acronym expansion, spelling normalisation, pronoun resolution etc.) would be needed to understand the fact. The idea was to make a good choice of which NLP techniques to work on. Unfortunately, many different techniques would be needed. In the event Tara's project never got as far as applying NLP techniques to search, but she has spent the last eight years working for Google in Mountain View, CA.

4.13 ∗The Information Retrieval Facility (IRF) in Vienna

In around 2006 I became a member of the Scientific Advisory Board of the IRF. Francisco Webber had created the IRF and complementary company Matrixware as part of his grand vision to improve

the capabilities of *professional search*, starting with patents. The vision was that he would create a very large scale computing laboratory and endow it with vast sets of patent data, after value-adding curation. Scientists from around the world would apply to the not-for-profit IRF to run research projects on the facilities. Scientific discoveries would be open sourced, and Matrixware would, like any other company, be able to exploit them while developing high-quality products and services to sell.

2006: Francisco Webber (Director, Information Retrieval Facility (IRF), Vienna) enjoying a bush barbeque on a visit to Canberra.

Francisco was a remarkable person of Austrian/Angolan heritage. He spoke multiple languages and was a qualified medical practitioner. He was a persuasive communicator of his ideas and plans.

He recruited a great team of scientists and IP professionals to the advisory board, and we gathered annually in Vienna (business class fares paid by the IRF) for a business meeting and conference. We became members of an Austrian verein (non-profit society). We stayed in top Vienna hotels and ate in places like the Hotel Imperial, where Adolf Hitler once worked. We also attended a concert in the MozartHaus, where Wolfgang Amadeus Mozart had lived for a short time in the 1780s.

Daniel Schreiber was treasurer of the IRF. He sometimes wore a T-shirt with the message: *Austria. No f___ing kangaroos!*. Three very capable and friendly people, Sylvia Thal, Marie-Pierre (MP) Garnier, and Babeth Piveteau organised meetings, conferences and travel for the IRF. They also organised my travel to represent the IRF at an Intellectual Property conference in Mestre, Venice. In Mestre I spent considerable time and emotional energy in defending myself against a barrage of email attacks – totally unjustified I hasten to assure you. On the day after the conference, Sylvia, MP, and Babeth dragged me away from this unpleasant activity, and persuaded me to accompany them on a walk around the old city of Venice which I hadn't visited before. They also advised me on the purchase of a present of Murano glass for Kathy.

IRF Travel Tribulations

Many things went wrong on my trips for the IRF. My first trip to Vienna had been booked by an Austrian travel agent who, without me noticing, had scheduled only 45 minutes between arrival at Sydney Domestic on Qantas to departure from International on Austrian Airlines (who closed their check-in 45 minutes before departure!) Of course I missed the check-in. Austrian Airlines were utterly unhelpful and quite unsympathetic. A very junior employee took pity on me and escorted me to some Qantas office and told them that it was their responsibility because their incoming flight

was late. Qantas ended up wearing the cost of the travel agent's bungle, and paid for a night's accommodation in Sydney, and bought me a business ticket on Austrian for the next day. I arrived only an hour or so before the meeting started but managed to stay awake.

2013: Babeth Piveteau was a key member of the IRF team. After the demise of IRF she moved to London, and having met again at the Information Online conference, we arranged to visit the newly completed, Renzo Piano-designed Shard[37] , near London Bridge. It's the UK's tallest building and as you can see, on a clear day the view would have been magnificent.

Having learned from the first IRF trip, I later insisted on longer lay overs and on flying Qantas to avoid the need to check in again in Sydney. My trip to Venice was to be on a new Airbus A380. Unfortunately, the aircraft scheduled to fly the QF1 service broke down. At that stage Qantas had very few A380s and had to wait until one returned from Los Angeles. Qantas booked me into a luxurious 5-star hotel for the night and, next morning, brought me back to the airport. While again waiting in the lounge, one of the staff approached me and handed me a complimentary First Class boarding pass. The resulting flight was the most comfortable I've ever experienced.

When we landed at Heathrow, a member of cabin crew accompanied me all the way to the baggage collection and ensured that my bag had arrived. Service indeed! I'd arrived too late for a connection to Venice, so they provided a luxury room at the airport. Amazingly on my eventual return to Canberra, there was a letter from Qantas apologizing and enclosing an $800 travel voucher to compensate me for the "inconvenience". Customer service far beyond the normal!

On another occasion I was driven from Vienna airport to the city at more than 160 km/hr in a taxi. Fortunately, when we had a small accident, it was at much lower speed – not much damage done, but half an hour wasted while the drivers sorted out the complex legal and insurance paperwork.

The worst IRF travel experience was when my wife Kathy joined me at the Imperial Riding School hotel after the IRF conference. We went for a walk and came back to find our three high-end laptops had been stolen. Most annoying! Particularly as Kathy had to present a talk at a conference in Salzburg in a couple of days time. We first scurried to change passwords, but the PC in the hotel's business centre was utterly useless, even after mastering the German keyboard and interface. You were limited to a maximum of 30 minutes and the connection was about 110 bits/sec! We went out to buy a Windows laptop, but found that they were only sold with a German version of Windows. We could buy an English version but would have to use the German interface and documentation to install it, and then also buy Office apps. The PC machines we were offered did not include an international warranty. We went to an Apple shop instead and bought a Mac laptop. Although it

[37] https://en.wikipedia.org/wiki/The_Shard

had a German key layout, we were able to switch the operating system and applications to English by ticking a single configuration option. And it had a world-wide warranty.

Enough of the negative travel stories!

*Professional Search

Francisco Webber's claim that *professional search* is a world apart from web search is absolutely true. Whether professional search is conducted in a commercial intellectual property department, by an academic researcher doing a systematic review, by an e-discovery paralegal, or by an investment adviser looking for opportunities to tempt HNWIs,[38], the process of searching is utterly different from that supported by Google or Bing. Full recall is critical, there is no silver-bullet answer and large amounts of time and money may be invested in each search – the IRF reported that the German company BASF allocated €30,000 per patent search. As we will see further on, IP (intellectual property) searches may determine decisions involving hundreds of millions of dollars.

One year at the IRF conference there were a pair of complementary half-day tutorials: IR for IP, and IP for IR. I attended the latter, and learned a lot about intellectual property. One of the tutors was Pierre Buffet, Executive Vice President of Questel, a vendor of Intellectual Property software and services. He and I had some interesting conversations and I gained insights into what extra capabilities would be needed in Funnelback to compete in the patent space – masses!

An IP professional from a Swiss agricultural chemical company posed a challenge for the IR researchers. He said that the breadth of applicability of many chemical patents was expressed in the form of "a chemical from among the list A, B, C, D, ... in combination with a chemical from among the list Z, Y, X, W, ...". Query languages provided by state-of-the-art patent retrieval systems inevitably produced thousands of false positives which had to be manually worked through. I contacted him later by email and asked whether he would be interested in funding CSIRO to work on this problem. Unfortunately, not!

One of the IP experts engaged to advise the IRF had been the head of the IP department at Akzo Nobel, a major chemical company. He recounted the story of Akzo Nobel's decision to manufacture a chemical previously only sold by a major competitor. His IP department had reviewed the patent protection for production of the chemical, and determined that Akzo Nobel was free to make and sell it. On that basis, they spent hundreds of millions of euros on building a factory. Of course, the competitor took them to court. Court cases in the USA and Europe cost hundreds of millions, but failed to fully resolve the issue. The factory continued to operate.

The Demise of the IRF

Unfortunately, Matrixware failed to generate revenue and the bank which had funded it and the IRF eventually pulled the plug.

4.13.1 *Retrieval by Textual Annotations

It had long been clear that a document could be retrieved on the basis of external anchor texts which referenced it, even without access to its content. Tom Rowlands's PhD addressed the generalisation of this concept to include other forms of external textual annotation. He looked at click-implied descriptions,[39] folksonomy tags,[40], and tweets containing URLs.[41]

Tom, Matt Adcock and I published a study[42] showing the benefit of using click-implied descriptions in search rankings for a number of P@NOPTIC customers.

[38] High Net Worth Individuals

[39] If a person submits a query q and clicks on a document d, then the text of q can be considered to be a description of d.

[40] Folksonomies are collections of tags applied to documents by users. At one stage they were proposed as the solution to providing high quality enterprise search, but there were never enough to be of significant value.

[41] The text of a tweet containing a URL may be considered as some sort of description of the document at that URL.

[42] https://david-hawking.net/pubs/hawking-rowlands-adcock-adcs2006.pdf

Tom also built a demonstration web search engine based on tweets.[43] It wasn't supposed to provide full coverage of the web, but instead to provide more rapid response to breaking events than a conventional search engine could. Tom's system repeatedly took in tweets from Twitter's garden hose feed,[44] dropping tweets without a URL. URLs in tweets are usually shortened, and the system needed to make another web access to find the actual target URL. At regular intervals a new batch of accumulated URLs was added to the collection and the oldest batch discarded. When someone searched for `cyclone` on Tom's demo it would be possible to find an article about a cyclone which had just hit a city, provided that someone had tweeted about it.

Tom found an organisation which had multiple types of annotation. It was the PowerHouse Museum in Sydney. Not only did it have anchor text and click data, but it also encouraged visitors to provide descriptions of exhibits on the web pages describing them. This was a sort of folksonomy, but unfortunately the number of tags was tiny. We believed that this was at least partly due to the fact that tags were anonymous.[45] In other non-anonymous tagging systems, you can apply your own tags and return to them later, returning to exhibits you've previously looked at. This provides greater incentive to tag.

Developing ideas from my joint work with Tom, I eventually built a two-level indexing and retrieval system based entirely on annotations. It was called ANNIE.[46]

4.13.2 *Document level security

One of the major challenges in enterprise search is that of providing reliable document level security. Documents in a collection may be subject to an access-control list (ACL) specifying which individuals or groups are permitted or not permitted to view it. We didn't want P@NOPTIC to implement its own parallel security scheme for checking user credentials against ACLs, out of fear that a future P@NOPTIC bug might expose confidential information to the wrong person.

For some organisations it was critical that access tests be performed at the time of the search – an employee changing roles should instantly lose access to documents restricted to their former role. The only viable way to safely satisfy these constraints was to use the operating system to test the visibility of every candidate search result to the current user.

Peter Bailey, Brett Matson, and I conducted a study identifying the costs of search-time document level security checking.[47] We found that it in some circumstances it could cause dramatic slowing down of search response. Access checks for an unprivileged user submitting a query with a very large results set would be very time consuming.

4.14 CSIRO – Constant State of Imminent Reorganisation

As previously mentioned, CSIRO projects came and went on an annual basis. Researchers wishing to remain active in their field of expertise needed to frequently repackage and rebrand their work as a new project. In the months before the spinoff, I had defined a new project called SEFA, Search Everything For Anything. The plan was to bring in other members of Ross Wilkinson's team and encompass types of search not addressed by the P@NOPTIC product. For example, video search, audio search, and effective search of email.

4.14.1 BHAGs and PeopleFinder

When Geoff Garrett became CEO of CSIRO in 2001, and had discussions with the Australian government, he feared that CSIRO was at significant risk of being split up, and/or having its funding

[43] `https://david-hawking.net/pubs/rowlands10.pdf`
[44] A small proportion of their firehose tweet feed.
[45] `https://david-hawking.net/pubs/rowlands_adcs08.pdf`
[46] See Page 155.
[47] `https://david-hawking.net/pubs/cikm127_bailey.pdf`

seriously cut. He reported that the government wanted CSIRO to address and solve problems of national importance. Internally, Geoff workshopped the idea of BHAGs – Big Hairy Audacious Goals. He also saw problems with the CSIRO's traditional silo structure, in which each CSIRO Division operated independently of others.

Over time, he took significant funding out of divisions and created Flagship Programs to address BHAGs of national importance, preferably with external co-investment. Example Flagships were *Wealth from Oceans* and *Water for a Healthy Country*. Flagships were required to bring in people from multiple divisions. In order for divisions to keep all of their staff employed, they had to contribute many of them to projects in the flagships.

When putting together the staff of a new or expanding flagship, CSIRO faced a challenge in identifying CSIRO people with the right skills. A senior researcher in the CSIRO ICT Centre, Peter Corke enrolled in a CSIRO leadership training course and, as a leadership project, defined a system for locating talent within the organisation. Geoff Garrett attended the project presentation and directed that such a system should be built, and that I should be involved in building it. The resulting system was initially called the *Capability Browser* but would eventually be known as *PeopleFinder*.

From previous experience with `P@NOPTIC Expert`, a CMIS expertise finder based on researcher web content,[48] I realised that that path was fraught with difficulties. Some researchers were angry that others were ranked more highly on key topics than they were. Some researchers had multiple internal profiles, some created without their knowledge. Some had no profile at all – some scientists believed that it was shameful self-promotion or a waste of time to put anything on the web – what mattered was publications in high-impact journals. Profiles were scattered in different places and had no consistent URL pattern. Most profiles were way out of date.

An initial kick-off meeting for the PeopleFinder project in 2006 involved around 30 people, but involvement dropped rapidly, and in the end the key players were Antony Stinziani (project manager), Catherina O'Leary (CSIRO Communications, Project Owner), James Dempsey (Java developer), and me (technical architect). I laid down some key principles which I regarded as essential for success:

1. Every CSIRO employee should have a PeopleFinder profile. (Not just scientists.)
2. No person should have more than one PeopleFinder profile.
3. The URLs of PeopleFinder profiles should conform to a simple pattern: e.g.

   ```
   .../PeopleFinder/IDENT.html
   ```

   where IDENT was the unique CSIRO identifier assigned to each employee.
4. As much as possible of the basic content of each profile should be automatically populated from authoritative sources. E.g. phone numbers from the internal directory, HR information from the HR database, project information from the projects database, and so on.
5. Employees would be given an interface to allow them to enter information about interests, skills, publications etc. and encouraged to do so.
6. Each item of information in the profile would be accompanied by a flag indicating whether the employee was willing to have that information published externally to CSIRO.
7. The PeopleFinder search interface and the interface for updating profiles should be developed by progressive refinement of prototypes.

CSIRO had not made a decisions on whether PeopleFinder profiles should be externally viewable. On the one hand, doing so might promote CSIRO's reputation for expertise, and foster opportunities for collaboration. On the other, it might expose us to poaching. If profiles became externally visible, I wanted employees to have the ability to suppress visibility of information they wanted kept confidential. For example, in unusual cases, they may have been involved in family violence or contested custody situations.

[48] See Section 4.5 on Page 57

I added the last item to the list because I knew how hard it is to specify a highly usable interface in one shot. I also realised that there was a significant chance that resources for the PeopleFinder project might suddenly evaporate. If that happened we would be better off with a not-fully-refined prototype than with a pile of design documents.

I was unable to persuade them of the merits of progressive refinement of prototypes, but when resources did evaporate, the system was working and seemed to be useful. Catherina O'Leary had done an good job with wireframe UI designs and James Dempsey had done a great job of implementing the system according to the first 6 principles.

The system was still operating when I left CSIRO in July 2008. Ideally we should have conducted a study of the purposes to which it was put, made a C-TEST file of representative queries and their right answers, and tuned the retrieval function, but I imagine that never happened.

4.14.2 The Cnawen Proposal

When setting up email aliases such as `panoptic@cmis.csiro.au` in the very early P@NOPTIC days, Francis Crimmins arranged to file all the mail and convert it to web format using the `hypermail` system. That webmail corpus was made searchable and proved very valuable to Panopticians in customer relationship management, knowledge archiving, and in finding solutions to technical issues.

For search purposes there were a few deficiencies in how `hypermail` worked, but despite this the P@NOPTIC mail archive was a very effective *corporate memory*. Having a centralised organisational memory like this seemed superior in many ways to relying on employees to retain and personally file what they considered to be useful. For a start, it could be comprehensive, and reliably backed up. It could facilitate onboarding of new staff, role handover, and legal discovery. It had the potential to support useful value-adds. Finally, it addressed the issue of employee privacy because personal email sent or received by employees would not be in the archive. I presented the idea of a generalised *Corporate Mail Manager* at a number of seminars I gave in the mid 2000s.

In early 2008, I was set the challenge of defining a commercialisable project which would involve as many of the members of Ross Wilkinson's Information Engineering team as possible. I was set a very tight deadline. I thought that there was a significant opportunity to do this by expanding the Corporate Mail Manager concept into something I called *CRACK!* I sent around *An Invitation to Get Cracking!* and organised a two-day workshop in Bowral for interested participants from Sydney and Canberra.

The outcome of the workshop was a change of name to *Cnawen* (meaning 'know') and a frenzy of enthusiastic activity. Carsten Friedrich, Alex Krumpholz, Peter Thew, Stephen Wan, Andrew Lampert, and Tom Rowlands were among the group who I remember as being particularly active. In an amazingly short time, they set up a web site, registered a trademark, and built a very impressive prototype bringing in tools for summarisation, timelines, visualisation, and relationship and commitment extraction. A professional sales brochure with screenshots from the prototype is reproduced in Appendix 11. Since the days of S@NITY, we'd learned a hell of a lot about how to do this sort of thing.

All this failed to persuade ICT Centre management but, for me, Cnawen was the most exciting project that never happened. ☺

4.15 Breast Cancer

At the very end of 2006, I was diagnosed with early stage breast cancer – yes, males can get breast cancer. My breast surgeon told me that he'd operated on more than 2000 women; I was the tenth man. He said that males always get mastectomies, since there are no cosmetic issues. On the positive side, with clear margins and cancer-free sentinel lymph nodes, that meant that I had no need of radiotherapy or chemotherapy. More than 15 years later, I realise I got off very lightly.

I delayed my surgery until Australia Day so that I could finish a paper I wanted to submit to the SIGIR conference. Of course, it was rejected. ☺

It was an unpleasant process being enmeshed in the medical system, with loads of invasive and claustrophobic tests. Given that my mother had died, aged 36, of the same cancer, there was a significant emotional shock, and a desire to take stock of life's direction, to re-assess priorities, and to spend time on what was most important. CSIRO and Funnelback people were very kind and sent cards, flowers and grapes, and even came to visit me during my very brief stay in hospital. Cécile Paris told me that she was very sad to add breast cancer to the things we had in common. HR people Anthea Roberts (now Anthea White) and Sarah Savage were very supportive during my recuperation.

2008: ICT Centre barbeque: Darrell Williamson, John Zic, Anthea Roberts

The 2007 ICT Centre conference was held at Olympic Park in Sydney a few months after my surgery. I strayed from the path of science in my talk, by relating my cancer experience and its effect on my attitude to life and work. I explained that I was even more motivated to do good science, and to minimise time spent on low-value distractions. I wanted to increase awareness of breast cancer, particularly male breast cancer, and to convey my sense that we CSIRO scientists should spend more time on actual science and less on the processes around it. I even swore in Dutch. It wasn't my intention to gain sympathy but after the session a remarkable number of my colleagues came up and hugged me.

4.16 Departing CSIRO

My frustrations with CSIRO deepened as time went on. The infuriating increase in bureacracy was, well, infuriating. The first annoyance was *effort logging*. Accurately recording time spent on externally funded projects was of course critical, but it didn't make sense for internal projects, and the effort logging system had major faults. If you worked a total of 76 hours in a week and recorded that, the initial system would scale it down to the nominal 38 hours. That was disastrous if you spent 19 of those 76 hours on a project for an external client, because the system said that you could only charge for 9.5. I also found that I worked hours which couldn't be accurately recorded because projects had dropped off the system or had not been entered.

2007 ICT Centre Conference at Olympic Park, Sydney. Me, Leila Alem, Alex Krumpholz. ***Photographer unknown.***

2008: Sarah Savage in the CS&IT building on the ANU campus.

I tried hard to accurately record times and installed a time logger on my laptop. I would diligently start it recording, but as soon as I started trying to solve a problem I'd completely forget the timer until I noticed it, still recording, the next day.

"Annual" performance assessment (APA) reviews became far more detailed and were conducted several times a year! This micro-management may have been effective in performance managing the unmotivated and the floundering, but for those of us driven to achieve and keen to get on with the actual work, it was a demotivating distraction.

A new requirement added around this time was that scientists in IT were expected to publish two conference papers and a journal article each year. This may be a reasonable expectation for an average over time but it became silly when translated into concrete goals in the APA. "What will be the titles of the three papers resulting from the research you haven't yet started?" "In what venue and in what month will they each be published?" This was in stark contrast to my CSIRO induction from Rhys Francis in 1998. He said that:

> In universities, academics are required to churn out papers, to publish or perish. In CSIRO it's not like that David. We want to know:
>
> - What problem did you work on?
> - Was it an important problem?
> - Does it matter?
> - Did you solve it?
> - Was it a good solution?
>
> If you can answer yes to all those, then we want you to publish in the best possible venue.

Project reviews became more onerous and more frequent. Plans became more detailed and progress on every item was rated on a rainbow of traffic light colours. Everything was supposed to be planned to the last detail.

At about this time, Geoff Garrett (CSIRO CEO) organised a round of meetings with scientists who'd had unusually high impact outside the organisation, to try to identify success factors which he could foster. I was flattered to be invited to the Canberra meeting. What I heard there was that every one of the big successes started as a departure from official plans – individual scientists and division chiefs had been given enough flexibility to follow up a lateral idea. To my astonishment and disappointment, the value of freedom and flexibility didn't make it into the official note-taker's report! ☹

Shortly prior to my departure, CSIRO introduced a matrix model in which scientists reported to at least two different leaders – a science leader, and project leaders. The science leader was automatically assigned according to the scientist's research area. On the project side, project proposals were put up to a planning team in Limestone Avenue, each with a spreadsheet listing the contributions of each scientist to be involved. Scientist S might report to Science Leader L and 30%, 50%, and 20% to Project Leaders P_1, P_2 and P_3. This model seemed to me likely to lead to confusion and conflict of priorities. In practice, making central decisions about projects was difficult. There were significant time gaps between the termination of old projects and decisions about the new ones.

The failure to approve the Cnawen project[49] was a disappointment to me. Furthermore, the ICT Centre external Science Review had rated Ross Wilkinson's Information Engineering group/theme/team at the highest available (world-class) level, but that failed to lead to plaudits and increased resources. Instead, our part of the ICT Centre made a major change of research direction which didn't align with my research interests.

I applied for five weeks leave to clear my head but was given approval for two lots of leave – two weeks and three days, a day in the middle to attend a one day workshop on the new research direction,[50] and two weeks and one day of leave.

Coupled with my distaste for frequent CSIRO reorganisations and for the approval or otherwise of projects by a small central committee[51] relying on duelling spreadsheets, I decided it was time to leave.

I applied for a full-time ARC professorial fellowship. My application was ranked in the top 5% but no fully funded professorial fellowships were awarded in that round. I also gave into pressure from colleagues working on Live Search (now Bing), and interviewed at Microsoft.

That was an interesting experience. Kathy and I were flown to Seattle on very short notice and I was driven to ten separate interviews and dinners in a chauffeured limousine. My interviewers included Harry Shum (CVP in charge of Bing), Ray Ozzie (then Microsoft CTO), and Satya Nadella (now Microsoft CEO). Several of the interviews were conducted on Good Friday.

Coincidentally, while in a hotel for those interviews, I received a phone call from the Chair of Funnelback, Steve Kirkby, offering me the position of Chief Scientist.

Microsoft offered me a Partner-level job based in Melbourne but I eventually turned it down because it would require me to move to Melbourne. (My wife Kathy's project was tied to Canberra,

[49] See Section 4.14.2 on Page 81.
[50] Oh, goodie!
[51] The Central Ministry of Planning?

and I considered her work on mental health more important than mine. I did eventually take up a position with Microsoft in 2013, based in Canberra.)

I left CSIRO for Funnelback in July 2008, after almost exactly ten years of service. Despite my accumulating frustrations, I owe CSIRO a considerable debt of gratitude. CSIRO gave me the opportunity to pursue a research career, and encouraged me down the path of research commercialisation.

Growing up in a small country town (Beechworth, Vic), I had been very aware of the great esteem in which CSIRO was held. People believed that CSIRO could be relied upon to give authoratitive answers to just about any scientific question. I was proud to become a scientist among its ranks.

Without CSIRO's confidence in me, I would have had a very different and less exciting career, Funnelback would never have existed, and this book would never have been written.

Chapter 5

CSIRO's P@NOPTIC Cottage Industry

This chapter describes the commercialisation activity around P@NOPTIC which occurred in parallel with the research described in the previous chapter.

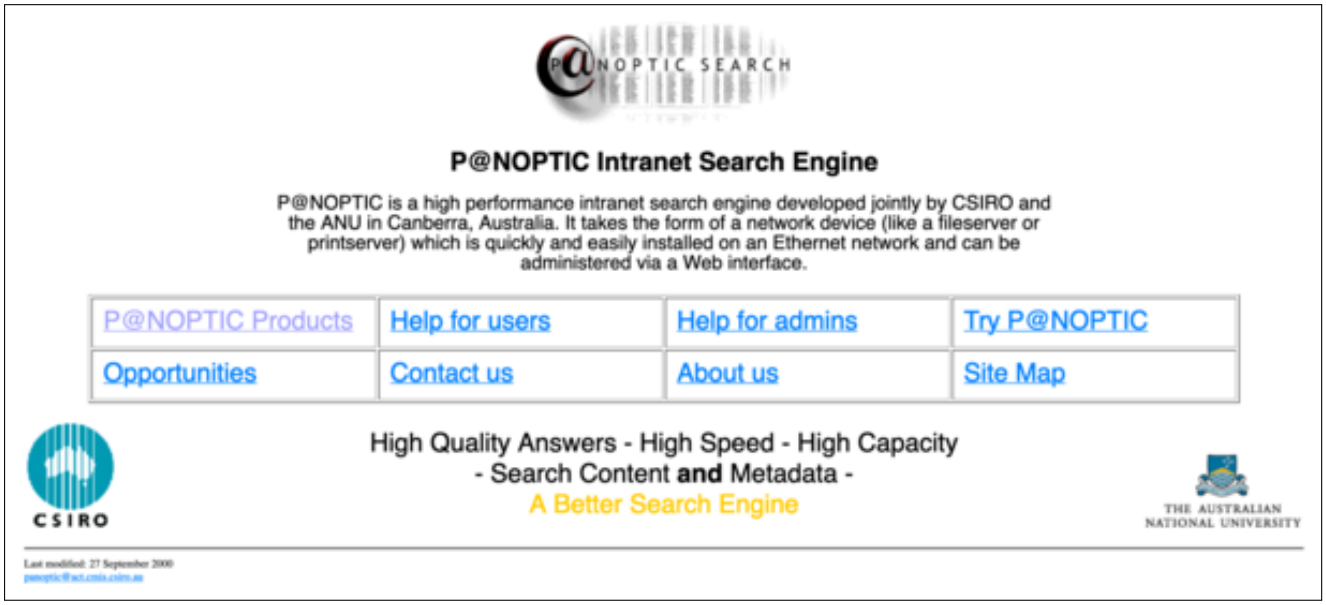
In around 2000, we geared up to run a small business within CSIRO. We obtained trademarks, registered the `panopticsearch.com` and `funnelback.com` internet domains, designed and printed brochures and business cards,[1] wrote content for commercial and technical parts of our web site, drafted P@NOPTIC licence agreements, and did our best to drum up trade.

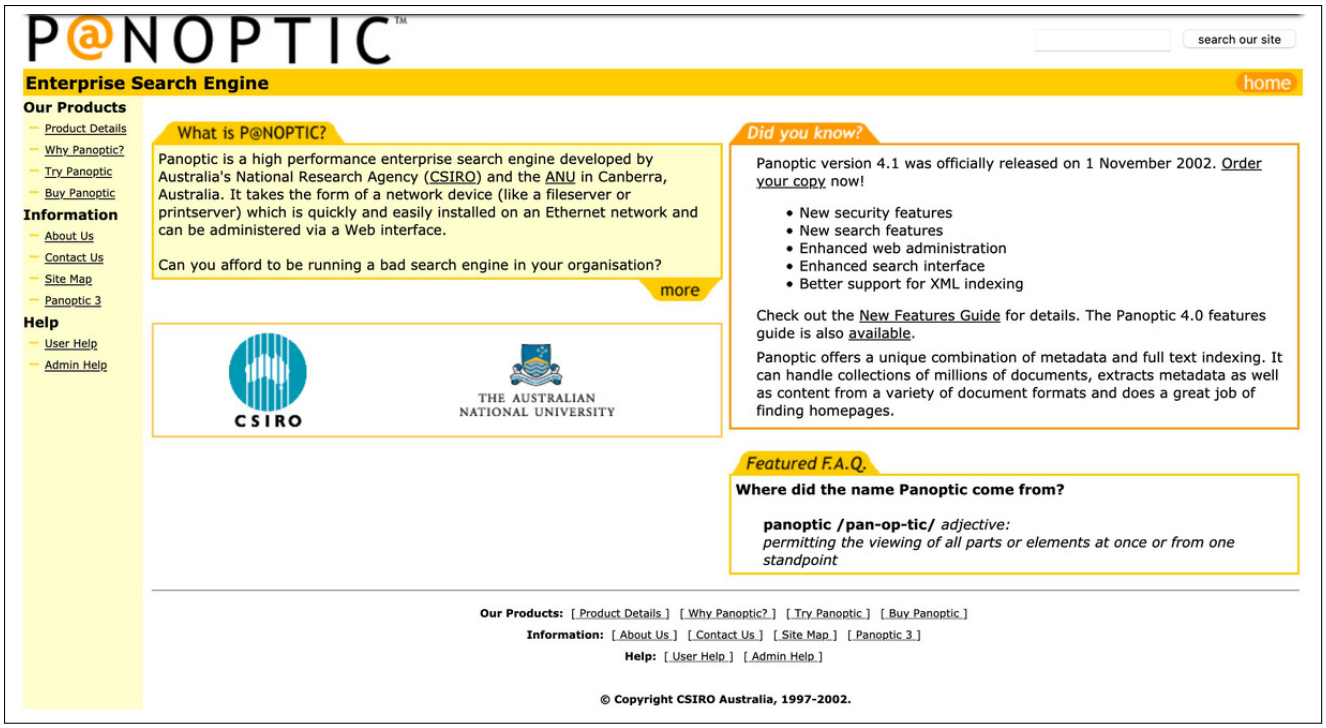
We started by approaching CSIRO Divisions and CSIRO as a whole, then universities and government agencies in the Canberra, Sydney, Wollongong area.

Ephemera from the Panoptic cottage industry as it operated in 2000. Top: Business cards.[2] Bottom: A roll of "powered-by-P@NOPTIC" stickers to be applied to customer search servers.

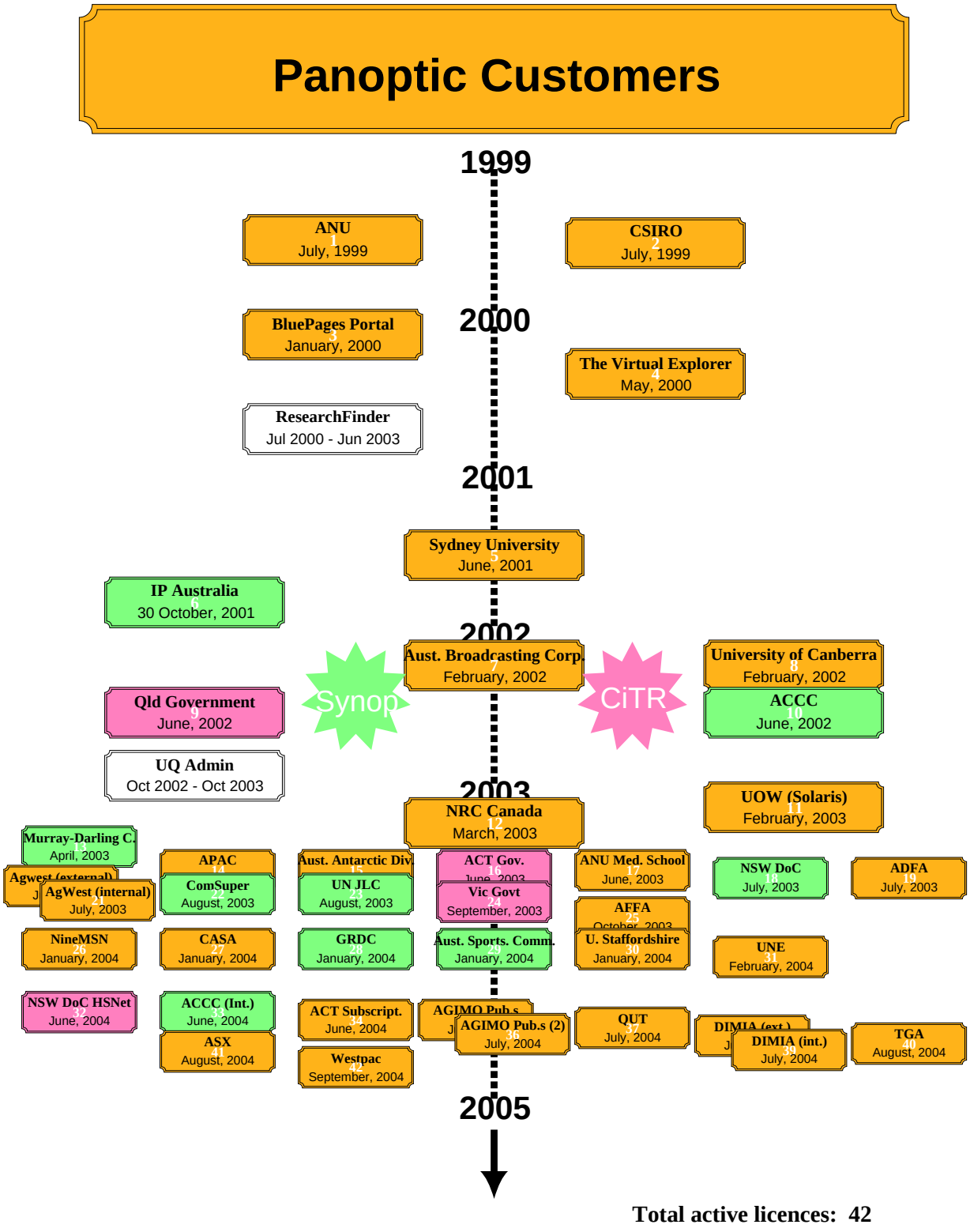
[1] See the appendices for the brochures.

[2] Note my commendable focus on containing costs – a business card shared by three of us.





Top: The P@NOPTIC commercial web site as of September 2000. Apart from the logos I fear I may have been responsible for the graphic design. Bottom: The same web site in early 2003, this time generated using Trystan Upstill's TrystanWeave system (see text).



My timeline of P@NOPTIC sign-ups to September 2004.[3] The first four didn't involve transfer of cash. ResearchFinder was the first commercially funded project. See the text for details. Synop and CiTR were licensed resellers of P@NOPTIC. Pink and green tablets show the customers of those resellers. The white tablets indicate projects which had terminated by 2004.

[3] Hand coded in PostScript.

The guiding philosophy of CSIRO's Enterprise Search Group

"License P@NOPTIC to as many customers as possible, use those installations to discover interesting search problems, do world-class research to solve them, implement the results in better search products, license them to even more customers, discover more interesting research problems, ..."

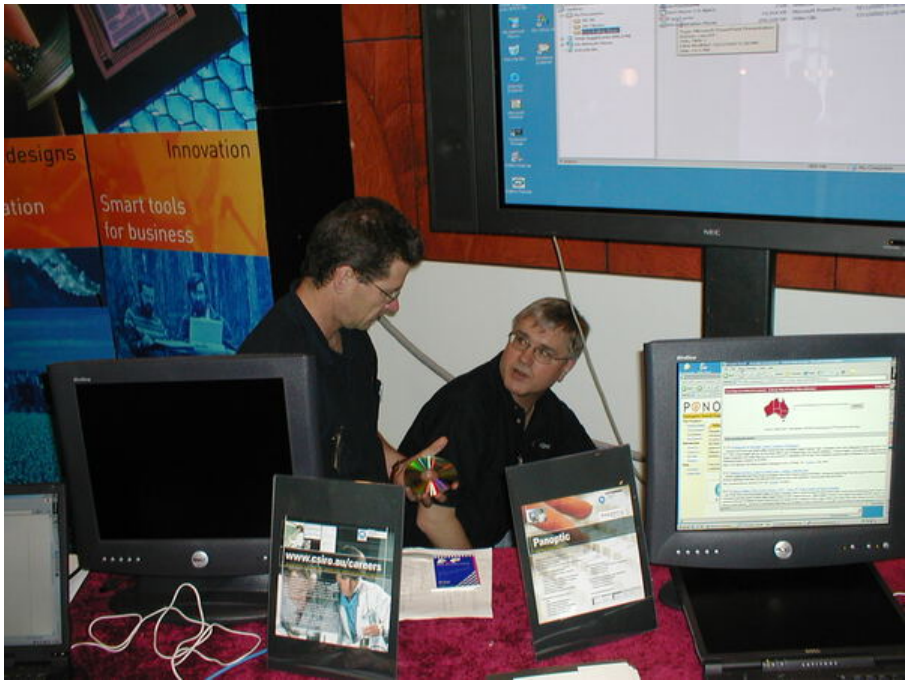
This was my enduring philosophy as leader of the CSIRO team working on P@NOPTIC. I was also keen that we learn everything we could about web search, through our own research and through the web search community and try to translate it into the intranet world.

There were a number of challenges in maintaining this aim which will come out as the story progresses. One of them was the major structural reorganisations which occurred within CSIRO each year. Not only did the names of organisational units change but the category names of the organisational units also changed – In a single year, projects became streams, SIMmeries became themes, and our division became a centre. Geoff Garrett, CSIRO CEO from 2001 to 2008, said that he'd been told by old-timers that CSIRO stood for "Constant State of Imminent ReOrganisation."

5.1 Eating Its Own Dogfood

Having persuaded ANU to operate a no-cost P@NOPTIC service to allow us to refine and showcase our technology, we naturally approached the other IP owner, CSIRO. At the time, the main `www.csiro.au` site offered a largely ineffective search service based on Microsoft IIS. CSIRO divisions operated their own web sites, generally without a search capability.

Discussions with CSIRO's ITS people were slow moving. People told us that although they could see the argument for "CSIRO eating its own dogfood", Bob Frater (then CSIRO's Deputy CEO) had very recently decided that CSIRO would exclusively "eat Microsoft dogfood." There were concerns about running a Linux server, and people were reluctant to use any of their budgets to buy the necessary hardware. Jonathan Potter, ITS General Manager was a supporter, but when he demanded action from his managers, one of them told him that action would be forthcoming once "additional resources" were provided. Finally, Peter Outteridge bought a server, and I installed P@NOPTIC on it, and trained a systems programmer Ben Tan.



Tom McGinness and Peter Outteridge setting up CSIRO's stand at an e-Government symposium at Parliament House.

The first CSIRO use of P@NOPTIC was actually by CMIS and the Division of Entomology. Entomology's IT Manager, Brennan Arrold, provided a testimonial for the earliest P@NOPTIC brochure. (See the appendices.) We were given an ancient Dell Optiplex[4] to run the services for both divisions.

There was no launch event for the whole-of-CSIRO search, and there were a number of problems in operating it. As noted on Page 70, a senior ITS person told me that P@NOPTIC should index only the Dublin Core metadata – it should ignore the content. ITS also insisted that CSIRO's external web content should not be included in the scope of the default internal search. A little further on you will see the adverse consequences of this decision.

P@NOPTIC was blamed by senior CSIRO people for a couple of instances of exposure of internal information. In one case the problem was misconfiguration of the web server. (See the panel on Page 43.) In another, the problem was a difference in definition of whether requests made by the crawler were "internal" or "external".

Then, a senior scientist from another division blasted the CSIRO P@NOPTIC service, at a high-level inter-divisional meeting, for being unable to find information that he urgently needed. As I did in the similar ANU case, I asked him for examples – what he searched for, and what he expected to find. Of the three examples, one didn't exist, another was not published by CSIRO, and the third was published on CSIRO's external site which, against my advice, was not included in the internal version of CSIRO search.

Finally, CSIRO bought a Content Management System (CMS) called Vignette, which was going to cause problems for search by publishing different content through the same URL. We needn't have worried – eighteen months after the purchase there were no externally visible CSIRO web pages created with Vignette.

We found that CSIRO IT Services was a difficult organisation to deal with, but it contained many helpful people. I've mentioned the help we received from Peter Outteridge. We were also grateful for support from Lani Cavanagh, a Web Communications Officer at CSIRO for more than 20 years. In around 2008 I worked closely with Catherina O'Leary from the CSIRO CommunicationsTeam on CSIRO PeopleFinder. See Section 4.14.1 on Page 79 for more details.

5.2 Research Finder

P@NOPTIC's first paying customer was the Commonwealth Department of Industry Science and Resources (DISR). John Thompson was running a project to make Australian research more accessible to potential partners and commercialisers. DISR wanted to make all commonwealth-funded research content on the web to be accessible via a single search interface: *Research Finder*. They wanted searchers to be able to restrict searches by state or territory, organisation type, or by FOR (Field Of Research) code.[5]

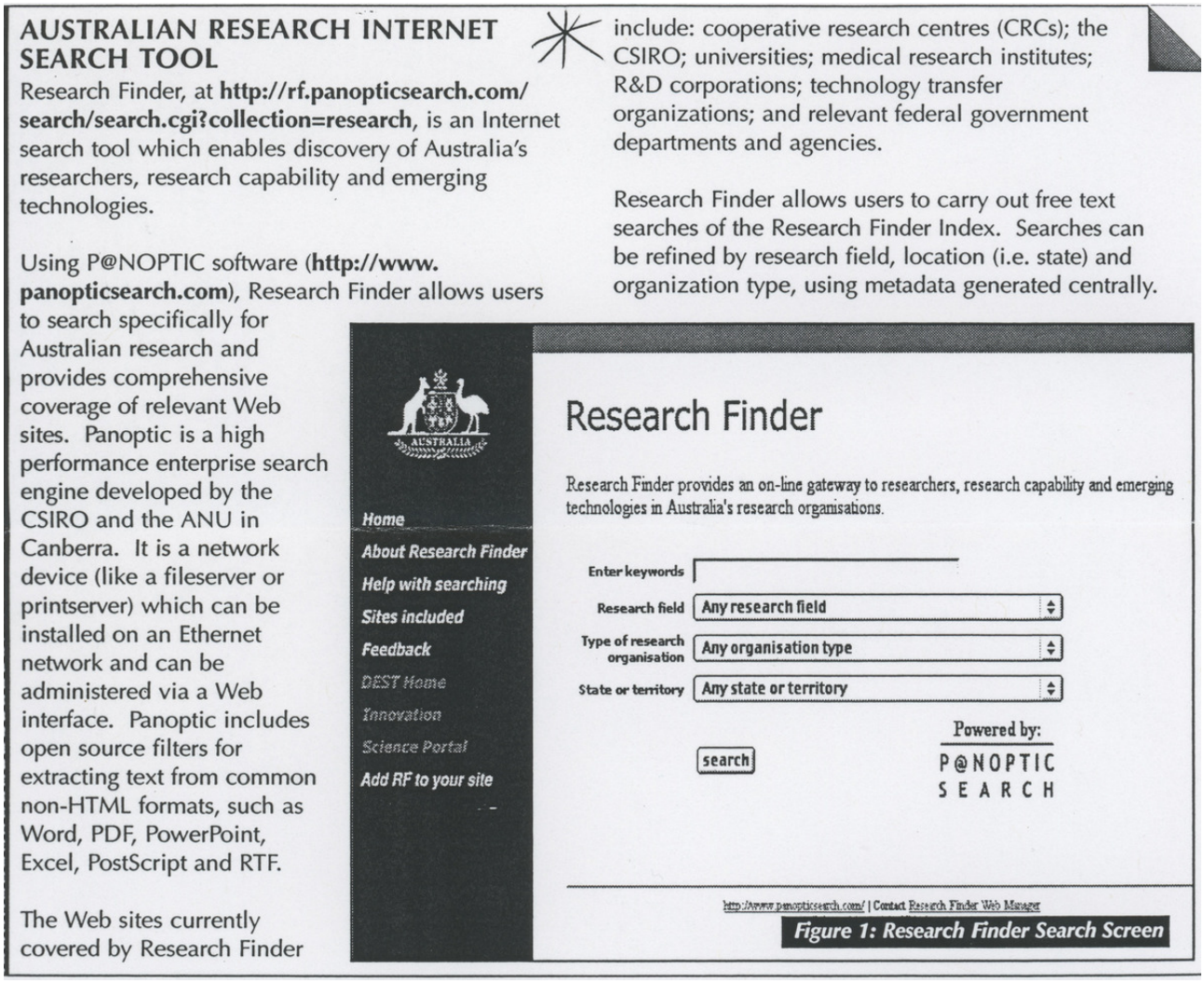
The trouble was that web pages and documents about commonwealth-funded research were scattered across universities, government agencies, and cooperative research centres. The web presence of many of these organisations contained large amounts of content which had nothing to do with research. DISR initially contracted with CSIRO to crawl and index content from a list of hundreds of different web domains. Francis Crimmins set up a server in his office to run the service.

Fran defined the overall service as three primary collections, each defined by a very long list of URL inclusion and exclusion patterns designed to gather all the research content and as little non-research as possible. Our scheduler interface allowed him to program non-overlapping updates of the primary collections, while Research Finder itself was a *meta-collection*[6] composed of the three primary collections. There were many crawling challenges and Fran established a Crawler Hall of Shame. The MIME type of one CRC home page was declared by the site to be of type AUDIO. ☹

[4] `https://en.wikipedia.org/wiki/Dell_OptiPlex`

[5] See `https://www.arc.gov.au/grants/grant-application/classification-codes-rfcd-seo-and-anzsic-codes` for details of FOR codes.

[6] When searching a meta-collection, the query was run across all the component primary collections and the results were merged into a single list.



AUSTRALIAN RESEARCH INTERNET SEARCH TOOL

Research Finder, at **http://rf.panopticsearch.com/search/search.cgi?collection=research**, is an Internet search tool which enables discovery of Australia's researchers, research capability and emerging technologies.

Using P@NOPTIC software (**http://www.panopticsearch.com**), Research Finder allows users to search specifically for Australian research and provides comprehensive coverage of relevant Web sites. Panoptic is a high performance enterprise search engine developed by the CSIRO and the ANU in Canberra. It is a network device (like a fileserver or printserver) which can be installed on an Ethernet network and can be administered via a Web interface. Panoptic includes open source filters for extracting text from common non-HTML formats, such as Word, PDF, PowerPoint, Excel, PostScript and RTF.

The Web sites currently covered by Research Finder include: cooperative research centres (CRCs); the CSIRO; universities; medical research institutes; R&D corporations; technology transfer organizations; and relevant federal government departments and agencies.

Research Finder allows users to carry out free text searches of the Research Finder Index. Searches can be refined by research field, location (i.e. state) and organization type, using metadata generated centrally.

November 2002: A report on Research Finder appearing in a magazine. I can't say for sure the name of the magazine as I have only a press clipping and no metadata. The top of the page is labeled Currents and there are many references to articles in *D-Lib magazine*.

Having built a basic searchable index, the next challenge was to assign external metadata to the millions of pages signalling FOR code, state or territory, and organisation type. DISR evaluated a classifier product from another local company but found that an excessive amount of effort would be required to generate the labeled data necessary to train the classifier. They were also sceptical that the results would be good enough.

To support this project, I was able to build into PADRE a hierarchical external metadata mechanism based on URL prefixes. For example, the line:

```
www.anu.edu.au/research/chem  x:ACT y:University z:Chemistry
```

would apply the specified x, y, and z metadata to every page in the index whose URL began with `www.anu.edu.au/research/chem`. Terms entered into the boxes in the screenshot in the clipping below would be prefixed with the appropriate x, y, or z field.

Nick Craswell built a web interface which allowed John Thompson and his colleague to manually specify the metadata which should apply to pages in the crawl. It cycled through all the servers and allowed the taggers to move in or out through the directories and sub-directories on that server. The prefix mechanism and the process by which metadata rules were applied allowed for great efficiency. Drop-down lists avoided the need to type the metadata values. John and his colleague tagged about three million pages in less than two person-weeks.

Example external metadata rules

```
www.anu.edu.au/ x:ACT y:University   # Applies to all pages unless over-ridden
www.anu.edu.au/NorthAustralia x:NT   # Over-ride for a subset of the content
www.anu.edu.au/research/chem z:Chemistry
www.anu.edu.au/research/chem/Organic z:OrganicChemistry  #Another over-ride
```

Research Finder was a great opportunity to test, refine and extend our technology. Once the service was properly established it ran very reliably. By the time Francis moved office and the machine had to be temporarily unplugged, it had been running continuously for more than a year. Since the crawling, indexing and searching placed sustained heavy loads on the machine, this was a good testament to the reliability of Dell hardware, Red Hat Linux ... and our own software.

Unfortunately, Research Finder ran into the same problem as the later Group-of-Eight expertise finder (discussed in Section 7.10 on Page 192.) Despite a large-scale effort to drive traffic, usage remained low, peaking for a couple of days after each publicity push and falling away again. DISR and the Go8 were keen to restrict search to Australian research content, but people searching for research were happy to put up with foreign pages among results from searches on Google or Bing. A query box accessing whole of web search was constantly accessible from their browser – no need to remember the name Research Finder and navigate to it.

After a couple of years, funding for the project dried up. In 2004 I contacted John Thompson to see if DISR would be interested in using P@NOPTIC on their intranet and external web site. He replied:

> ... DEST is heavily into Microsoft for its system and desktop software. DEST is currently redeveloping its Intranet and Internet using Microsoft Content Management Server but I do not know what decisions have been made already about site search capability. ...

5.3 The P@NOPTIC Search Appliance

Francis Crimmins joined CSIRO's P@NOPTIC team at the end of 1999. He had completed a Masters degree at Dublin City University which involved building a metasearcher.[7] He came with a strong reference from his supervisor Alan Smeaton, who was well known to me and Ross, from his participation in TREC.

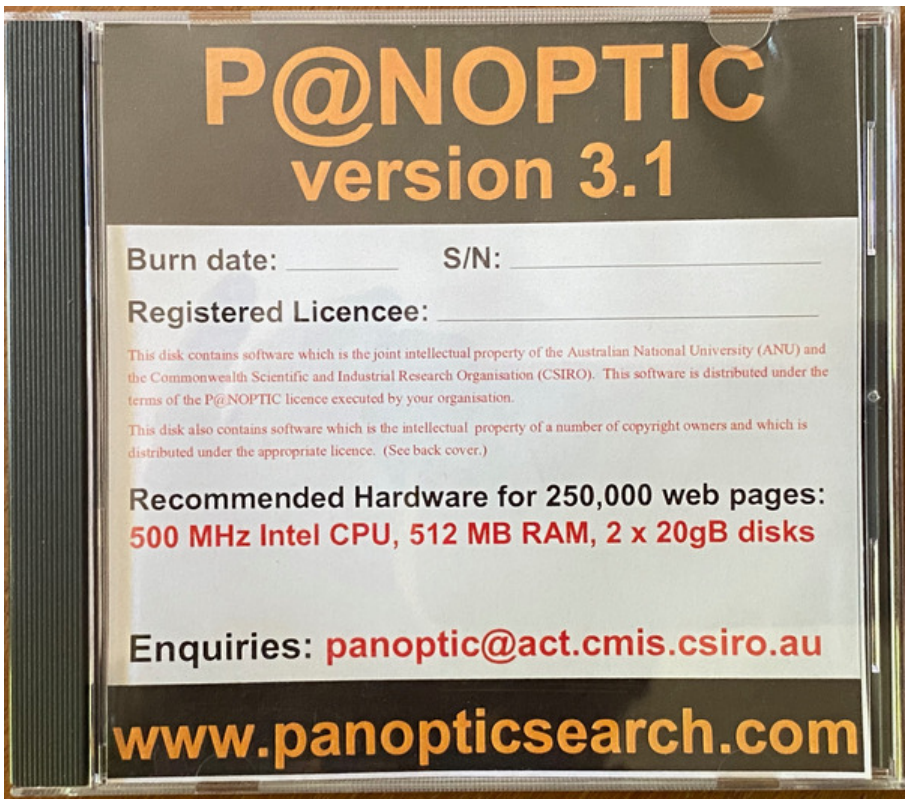
Fran suggested that we should build a search appliance based on P@NOPTIC – like a fileserver, printserver, or network router – a black box device which sits on the network and provides search services. I thought that this was a brilliant idea, since we could choose the operating system we wanted to work with and never have to worry about whether our code would run on all the different versions of all the different operating systems and hardware which might be run by customers.

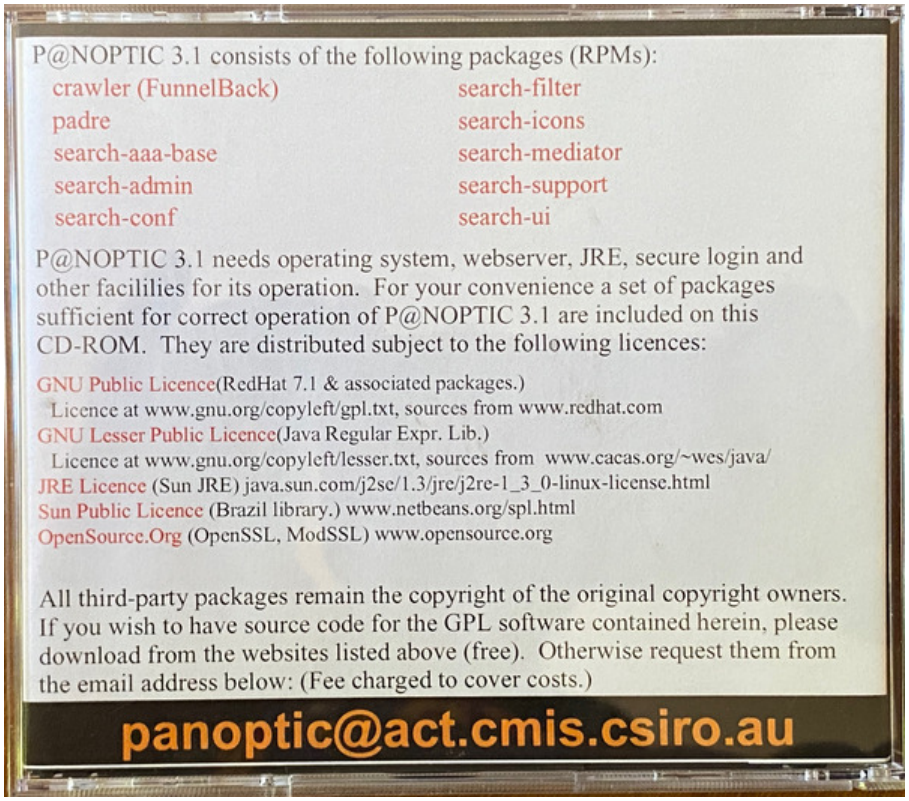
We gave thought to having P@NOPTIC-branded server casings made, and looked at products from Cobalt Networks.[8] We just didn't ever have the budget or marketing muscle to pull that off.

Fran set about creating a CD-ROM with a cut-down version of Red Hat Linux plus the P@NOPTIC packages and those for essential services like Apache webserver. He also started work on writing a crawler in Java, using the Heydon and Najork paper on the Mercator crawler as a guide.

We did indeed create a search appliance based on a Dell rack-mount server. Jie Gao from USyd IT Services had used P@NOPTIC at ANU and was enthusiastic. The University of Sydney signed up and, in June 2001, Ian Mathieson[9] and I drove to Sydney with a P@NOPTIC appliance in the back of the CMIS station wagon. Within half an hour or so, the USyd service was on the air and crawling content.

[7] A search engine which generates results by querying a number of other search engines and aggregating the results.
[8] https://en.wikipedia.org/wiki/Cobalt_Networks
[9] A member of Ross Wilkinson's Melbourne group who spent a few months with us in Canberra.





An early version of the P@NOPTIC CD-ROM.

Unfortunately, most other organisations refused to accept a P@NOPTIC appliance on their networks. "We're a Microsoft shop – we only run Windows NT", or, "Only HP hardware is allowed on our network." Naturally, these sites were running other appliances like routers, fileservers etc. which ran a different operating system on different hardware, but P@NOPTIC didn't have the credentials to push through. Google did have, and many otherwise-restrictive organisations were happy to install a Google Search Appliance when it was released.

2011: Institutional Web Managers Workshop, Reading UK. A Google representative shows off a Mini version of the Google Search Appliance.

Later, when Funnelback approached a large Commonwealth department about the possibility of using Funnelback on their intranet and external web sites, they responded along the lines of,

> Oh yes, search is really important problem, but we haven't got time to look at it seriously at the moment. We're just going to buy half a dozen Google search appliances[10] to tide us over in the interim. [Apparently with no competitive tender process.]

5.4 Pricing P@NOPTIC

P@NOPTIC was aimed at the *enterprise search* market. We wanted to help employees and other stakeholders to reliably find the information they needed to do their jobs effectively, wherever that information was located. There were a number of "industry" research studies which attempted to quantify the cost of ineffective enterprise search. A report by Susan Feldman for IDC corporation found that millions of dollars were wasted at a major airline because knowledge workers were unable to find required internal information. She highlighted the prevalence of cases where employees recreated information which already existed. On the other hand, there was a widespread belief that enterprise search projects failed to deliver value.

The biggest players in the enterprise search space were Verity, Autonomy, Fulcrum, OpenText, and later FAST Search and Transfer (which originated in the Masters thesis of Knut Magne Risvik at the Norwegian University of Science and Technology in Trondheim). Autonomy and Verity were known to charge millions of dollars for enterprise search projects, which typically included large allowances for consultants to install, configure and tune the basic installation.

One vision of ours was to create a search product which could plug into an organisation's network, and "just work". Because we were a research organisation we wanted to minimise the time our research staff needed to spend installing, configuring, and tuning. We also aimed to achieve a large volume of sales by setting an affordable pricing structure.

A second vision was to provide "bureau service" search, later known as "hosted search" or "search as a service". This was only planned to apply to search of externally visible content. Many organisations who these days probably keep confidential information in "the cloud" let us know in no uncertain terms that there was no way they would consider allowing their confidential information outside their own firewalled networks.

I was told to seek help from CMIS business development people on setting pricing. They asked me to prepare a draft schedule for them to comment on. My draft appears in Appendix 11. The principles behind the pricing structure are on Page 251. The outcome was, "Could be a bit high. Could be a bit low!".

[10] At high cost!

My pricing model for the bureau service was an attempt to model both the cost of us providing the service and the value to the customer. It was a formula based on number of documents, frequency of updates, and volume of queries. There was immediate push back from customers who said that having a charging formula based on usage (like telephones, internet, electricity and water) was unacceptable because it would not allow them to budget.

At this stage, in the very early 2000s, Google had not yet developed their Search Appliance but they were providing a version of hosted search, based on slicing their whole of web index.

5.5 Benchmarking P@NOPTIC Against the Competition

In the market generally, and sometimes in specific competitive opportunities, P@NOPTIC needed to compete against both expensive commercial search engines and free open-source alternatives.

We were able to demonstrate advantages over externally visible competitors using side-by-side comparisons, which also demonstrated how easy it was to install and configure P@NOPTIC. However, demonstrating superiority over open-source competitors on internal projects was not so easy. To establish that we were not wasting our time, CSIRO contracted with Paul Thomas, then an ANU PhD student to benchmark P@NOPTIC against the open-source crawler `nutch` and the open-source retrieval system `lucene`.

The comparison was favourable to P@NOPTIC. Later experience showed that organisations which chose open source solutions tended to be those who had application development teams capable of getting in and modifying the source code or integrating it into their applications.

5.6 A P@NOPTIC Peg in a CSIRO Hole

Focus, focus, focus!

Geoff Garrett had announced several key principles which he claimed would lead to CSIRO success:

1. Service from Science.
2. One CSIRO.
3. Partner or Perish.
4. Focus, focus, focus!
5. Look out!!!
6. Go for Growth.

I focused on "focus, focus, focus!" Later, at a party to celebrate the Funnelback spinoff, Geoff described me as the monomaniac inevitably to be found behind a successful project.

When I joined CSIRO I expected that Australia's leading science organisation would suit me down to the ground, by providing a wonderful opportunity to try to do world-class research and apply it for the benefit of Australia and the world. I was super enthusiastic. CSIRO's government-imposed requirement to generate external earnings caused me a little concern, however. It was said that the government applied the external earnings lever to ensure that CSIRO conducted relevant and useful research – research is by definition useful if people pay you to do it. I knew that there was valuable research that no-one would pay you to do, and that people would fund some types of research which had dubious scientific merit. Nearing the end of a financial year, CSIRO staff used to joke about lamington drives and chook raffles.

The standard way external earnings were generated across CSIRO's many divisions was through contract research conducted for large companies and government agencies. In agriculture, environment, mining etc. organisations could fairly easily be found to pay top dollar for CSIRO's outstanding expertise. In our Mathematical and Information Sciences (CMIS) Division, some customers were also to be found for statistical consulting. Unfortunately, in the Information Technology area, there

were very few organisations with either the budget or the inclination to fund scientific research. There was a lot of credible private sector competition for IT consulting projects – think IBM, KPMG etc. Many requests for IT consulting which CMIS did receive were for straightforward projects, not involving any scientific research, and easily carried out by the private sector.

I've already written about my early conclusion that a better approach was to license software which would (hopefully) improve the efficiency and productivity of Australian organisations while generating the necessary external earnings. See the panel on Page 89. If successful, this approach would hit the trifecta of research, impact and earnings. The cost to customers would be affordable because the cost of the R&D would be spread over many of them. For us, ongoing streams of licence revenue would smooth out cash flow across multiple years and avoid feasts and famines.

Unfortunately, there was a widely held belief across CSIRO that CSIRO should not be selling software. Indeed, I once overheard a senior CSIRO person telling someone from the Australian Research Council that, "CSIRO doesn't do IT research. That's best left to the private sector." Down the track I heard a senior CMIS person say publicly that CSIRO does not produce software for licence. This struck me as odd since that person had signed quite a number of customer contracts for P@NOPTIC.

In 2005 and 2006, the CSIRO ICT Centre organised an annual conference at Luna Park in Sydney. The next page shows my attempt to humourously relate images from the venue to CSIRO principles.

CSIRO employed a number of Business Development Managers. Because they were employed to drum up contracts for CSIRO research, they were not geared up for selling software licences. An exception to this was Trieu Hoang, who became involved in P@NOPTIC in around 2001.

In 2002, being in a senior CMIS manager's office for a meeting unrelated to P@NOPTIC, I was asked to stay behind and then given a severe dressing down, along the lines that I had created a massive liability for CSIRO by distributing software – albeit through licences that that manager had signed. Since we were not yet pulling our weight on the external earnings front, I was also accused of "featherbedding" on my colleagues.

2006 Luna Park ICT Centre Annual Conference. Bob Williamson (CEO of NICTA) explains something to Alex Zelinsky (Director, CSIRO ICT Centre) with typical intensity.

Service from Science

Go for Growth

Focus focus focus ..... aye aye Cap'n Zelinsky!

Partner or Perish

Look Out!

One CSIRO

CSIRO will own all the IP

My boss Ross

My approach to things was very frequently at odds with the way CSIRO worked. I chafed at the frequent re-organisations and changes of focus across the Division and later Centre.
Ross Wilkinson was my boss, or grand-boss, throughout my ten years in CSIRO. To the CSIRO leadership he supported my team's activities while to me he gently attempted to guide me along CSIRO's preferred path. He had a most unenviable job as the [vegetarian] filling between the leadership and the workers. He had to absorb my occasional expressions of dissatisfaction and frustration.
Ross and I have very different styles of working. Ross is an abstract thinker, keen on intellectual discussion leading to consensus, and willing to accommodate change. I remember a meeting of more than a dozen people whose teams had just been amalgamated into a group/project/theme under Ross. Their areas included intranet search, summarisation, grammars, video annotation, and personalisation of documents. The goal of the discussion was to see whether we could identify a project that we could all work on together. I found it a bit frustrating – I firmly believed that my team already had a project worth working on from scientific, community impact and commercial points of view, and I wanted to nail it.
The CMIS Melbourne lab was first moved away from the CBD (and its academic collaborators in RMIT and the University of Melbourne, as well as the most likely industrial partners) to the CSIRO campus next to Monash University in Clayton. Possibilities for collaboration with CSIRO's battery factory and the food science and nutrition division seemed to me rather limited. The move entailed a lengthy commute for Ross and his team. Several members of the team left CSIRO (either temporarily or permanently) rather than move to Clayton.
Later, after the Information Sciences people within CMIS had become part of the new CSIRO ICT Center, the decision was made to close the Melbourne laboratory. Ross stayed in his leadership role but was required to spend two or three days a week in Canberra. This pattern of life is tiring, stressful and unpleasant. I'm surprised that Ross was willing and able to maintain the routine for as long as he did.
I am very grateful to Ross for hiring me in the first place and tolerating my stubborn, single-minded pursuit of the P@NOPTIC goal. Although I wouldn't do much differently if I had my time over, I do regret any extra burdens I placed on him.

5.6.1 The 2002 Pipeline Review

Trieu Hoang at the Pipeline Review, on CSIRO's North Ryde campus.

Late in 2002, I received an email from someone I'd never heard of asking me to attend a pipeline review, whatever that was. I assumed it was spam, and laughingly reported it to Ross Wilkinson. He told me that it was actually very important. CSIRO's corporate commercialisation group had decided (or been informed by someone) that P@NOPTIC had reached the end of the CSIRO R&D pipeline – research an area, try to prove value, decide whether to sell it, spin it out or abandon it. CSIRO has to be prepared to "fail fast", I was told.
Trieu Hoang and I travelled to Sydney to present before the commercialisation experts. I did my best to argue the merits of my virtuous cycle model, but no-one seemed to be much interested, except for Stuart Beil who I'd never previously met. He got excited and said, "These guys have got a product, they've got customers, and they've got revenue!" Lo and behold, Stuart eventually gave up his role in CSIRO central and had himself transferred to CMIS, principally to work with us.

5.6.2 The 2003 CMIS Review

In early 2003 CMIS leadership forecast a serious budgetary shortfall. All scientists were required to submit a document arguing the value of their project and the case for their own continued employment. Scientists would be ranked in order of the judged merit of their cases, and the bottom fourteen would be given their marching orders. I didn't like this process as it had a serious negative effect on everyone's morale. Personally, if I were one of the fourteen I would rather be able to believe that the process was totally random. ☺

On the morning that outcomes were to be announced, our pigeonholes contained literature to help us have a more positive attitude to possible demotion or termination. Everyone was told to be in their offices in case the Chief wished to telephone to advise them that their career in CMIS was over. Fourteen people received such calls and were instructed to cease work and to wait to see if there were positions for them in other CSIRO divisions.

For us, the outcome of the process (in around April 2003) was mixed. My team was recognized as possessing scientific talent, but said to be pursuing a business model doomed to failure. We were downsized, and told to switch our strategy. Instead of licensing our software we were encouraged to earn money by upskilling Australian industry and government through seminars, training and consulting. The pressure on us was intense, and one of my team pleaded with me for a major change of direction to align more with the orthodox way of doing things.

Ironically, between that April outcome and June 30, a small rush of P@NOPTIC customers signed contracts and we beat our external earnings target by about 50%. Looking back, our revenue had been growing at a rate of above 80% per year, compounding from a very low base. Now, we were getting somewhere!

2003: Francis Crimmins, furiously debugging the crawler in his room at the Las Vegas Hilton.

Also that year, P@NOPTIC won an Editor's Choice award for best enterprise search engine from Network Computing magazine in the United States,[11] and I was awarded an honorary doctorate from the University of Neuchâtel, Switzerland for my contributions to the objective evaluation of web search. Francis and I flew to Las Vegas for the Network Computing Awards Dinner. Unfortunately, problems had been experienced with a critically important crawl, and Fran and I spent much of our Las Vegas time poring over Java code.

[11] Francis Crimmins saw the opportunity and provided the magazine with all the required information.

Newspaper advertising. Left: Computing section, Canberra Times, 26 May 2003. Right: The Age, 09 September 2003.

Despite the turnaround in revenue fortunes, I followed instructions and organised a series of *Search and Searchability* seminars in Canberra, Sydney, Melbourne, Adelaide, Brisbane and Perth. I explained how search engines work, how to evaluate the quality of a search engine, and how to publish information so as to facilitate effective search. The latter theme (search engine optimisation, SEO) soon became a significant industry. SEO services began to be provided by many organisations, some more scrupulous than others.

Some of the S&S seminars were very well attended – around 90 in Perth and more than 130 across two sessions in Canberra. Of course, it was made clear that CSIRO had its own product in the intranet search area, and CSIRO staff placed P@NOPTIC banners and brochures around the venues.

Over time, I presented similar talks at industry events around the country, and later in London, Edinburgh, and Milton Keynes. I had the good fortune to present at Australia House, Westminster Abbey,[12] the English Speaking Union, and on the Royal Mile in Edinburgh.

5.7 Whole of Government Search

My Search & Searchability seminars delivered a small harvest of opportunities for P@NOPTIC. After the first one at the CSIRO Discovery Centre in Canberra, I was approached by Julie Murphy and Glenys Gould from the Australian Government Information Management Office (AGIMO).[13]

They said that AGIMO was contemplating embarking on a project to make government online publications more accessible. They explained that the Office of the Government Printer had been shut down and agencies had been encouraged to publish information via the web. The trouble was that, with hundreds of government web sites, the public would need a portal by which to be able to access publications across the whole set of sites. Perhaps this could be achieved using a search service like P@NOPTIC.

Progress on the Publications project was slow, but in the meantime AGIMO engaged with us on other projects. In May 2004, AGIMO contracted with us to conduct a survey of federal government web content, and to write a report. The report was very comprehensive, identifying all the government web servers, and classifying their content. How many sites? How many pages? Distribution of pages? Distribution of types of content? Distribution of page sizes? Prevalence of metadata tags? How many pages have no Dublin Core tags? (Policy was that all of them should have DC metadata, but less than half had any.)

[12] In their conference centre, not from the pulpit. ☺

[13] It may have been NOIE (National Office for the Information Economy) at the time, but they soon changed to AGIMO.

2006: Glenys Gould of AGIMO and Brett Matson.

Apart from Julie and Glenys, other AGIMO people involved included Eric Davis, David Doherty, and Peter Alexander. On the P@NOPTIC side Tom Rowlands had joined Francis Crimmins and me on the project.

There were several extensions to the original contract, and one of them was to index all the crawled pages and provide a search interface for evaluation by AGIMO.

This must have been satisfactory because we were then awarded a contract to run a whole-of-government search service, which included not only the federal government web sites but those of state and territory governments as well. The acceptance certificate was dated February 2005. This was a major project for P@NOPTIC, and later for Funnelback, and provided a substantial part of our revenue. In 2005, P@NOPTIC's external earnings comfortably exceeded a million dollars, four times our target. This was due in no small measure due to Stuart Beil's ability to set appropriate prices and to close deals. There is a definite skill to being able to close deals, and I didn't have it.

The rapid rise in P@NOPTIC fortunes started the move for a P@NOPTIC spinoff, which was no doubt Stuart's plan all along. The spinoff story is recounted in Chapter 6, post-spinoff projects for AGIMO in Section 6.3 on Page 135.

5.7.1 Media Alerts

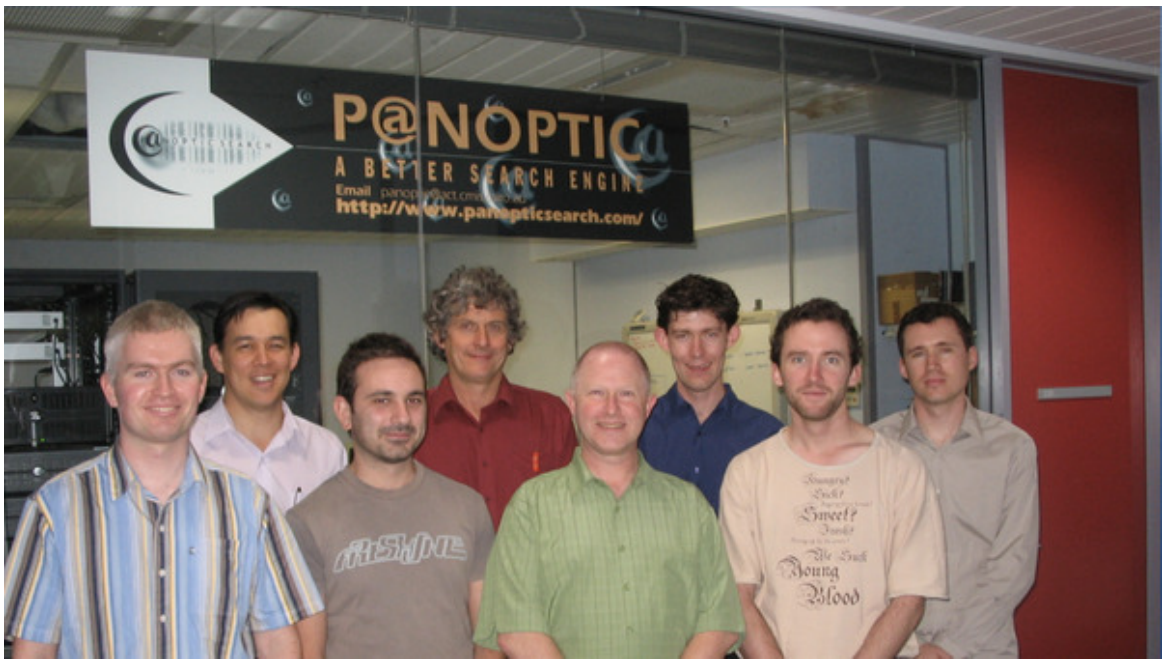
Another AGIMO project was to create and operate a media release alerting service. Francis Crimmins built a mechanism which would crawl a list of media-release sites, detect the releases which had been added since the previous crawl and index the changed content. He also built an interface by which members of the public could register and maintain standing queries along with the email addresses to which alerts should be sent.

After each update, an AGIMO person, usually Glenys Gould, would vet the new releases. Once approved, all the standing queries would be run and alerts sent for each new release which matched a standing query.

Francis anounces the media alerts demonstration.

```
From: Francis Crimmins [mailto:Francis.Crimmins@csiro.au]
Sent: Monday, 6 December 2004 3:46 PM
To: Murphy, Julie; Gould, Glenys
Cc: CSIRO Panoptic
Subject: Demonstration Media Release Search/Alert Service

Hi Julie, Glenys:

We've set up a demonstration search/alert service on government media
releases which you might like to try out:

http://bandicoot.panopticsearch.com/search/notify.cgi?collection=media

You can sign up for separate subscriptions for topics you are interested
in (e.g. "free trade agreement", "biometrics" etc.) and it will email
you an initial set of results.

The collection is scheduled to update on Wednesday. If new results have
appeared on your topic(s) then the system will email them to you. The
form above also links directly to the search service for normal queries.

The current demo is based on crawling to a distance of "2" from AGIMO's
media release site list:

    http://media.australia.gov.au/medsites.php
```

January 2005: Celebration of signing the AGIMO govsearch contract. Clockwise from front left: Francis Crimmins, Kit Chow, Stuart Beil, Alex Zelinsky, Tom Rowlands, Brett Matson, Trieu Hoang, Peter Thew, Gautam Tendulkar.



2005. CSIRO P@NOPTIC team outside the server room on the top floor of the CSIRO wing of ANU's CS&IT Building. Back row: Stuart Beil, me, Matt Sheppard, Brett Matson. Front row: Francis Crimmins, George Ferizis, Peter Thew, Tom Rowlands. *Photo: Sarah Savage?*

5.8 P@NOPTIC People

At the end of 1999, the CSIRO P@NOPTIC team comprised me and Francis Crimmins. The above photo was taken in the year leading up to the spinoff (2005) and shows eight members of the team. In the intervening period, Nick Craswell worked with us for about three years. Ian Mathieson, normally part of Ross Wilkinson's team in Melbourne spent several months working with us in Canberra, mostly assisting Francis with the crawler. Walter Stein, who gained a national record of 98 metres[14] in constant-weight freediving at Vertical Blue 2009, worked with us on an electronic payment system for my Search & Searchability seminars. Trystan Upstill, who was an ANU PhD student under the supervision of Nick Craswell and me, worked part time for us, creating a system (TrystanWeave) for generating our commercial and technical web sites. On completing his PhD, Trystan turned down our offer of employment and joined Google instead. Over time, he became head of Google's Search Quality team, then ran Google News. He is now Vice President of Engineering, in charge of Android R&D.

I've already mentioned the role of Trieu Hoang in providing business development and sales support in the first couple of years. Kit Chow was our legal adviser. She modified CSIRO contracts to serve the needs of P@NOPTIC licences, and assisted us greatly in the contractual situations which arose. This was in an era when government agencies outsourced their legal work to large law firms, providing an incentive, it seemed to me, for those law firms to "make work."

One government agency rang me at about 2am[15] to say that they would like to take up our offer to provide search for their agency. Kit sent them the contract and, after a very long delay, their outsourced lawyer contacted her to argue over terms and conditions in the contract. That lawyer claimed that the contract was unprofessional, embarrassing to CSIRO, and unworthy of the most junior of lawyers. It needed to be totally redrafted. Kit responded that the lawyer should take that up with their boss since he had drafted the contract for CSIRO!

[14] This means that, without an air supply, he dived to a depth of 98 metres and returned to the surface on a single breath.
[15] I was attending a conference in the UK and hadn't silenced my phone.

2004: Trystan Upstill at the 2004 SIGIR conference in Sheffield, England. (Hence the spoon sculpture.)

Francis Crimmins became my first recruit in late 1999 after arriving from Dublin, following his Australian partner Frances Duff. Confusingly, when they married she became Frances Crimmins. While working on the P@NOPTIC CD-ROMs and building the Funnelback crawler from scratch (as previously noted), Francis (Fran) was very alive to the commercial scene. He worked on scoping the size of the potential Funnelback market. By analysing a number of crawls of government and commercial domains in Australia, he could work out the number of organisations who had enough web content to justify running a site search. He could also tell which of those organisations operated a search engine, and what type of engine it was. This was invaluable information, because some of our competitors at the time performed very poorly. This gave us much better market data than we could get from consultants who were later hired to make business plans!

Fran was one of the three founders of Funnelback Pty Ltd and, soon after, became its R&D Manager. My role, initially part-time and later full-time, in Funnelback was that of Chief Scientist. I was careful not to undermine the R&D Manager, who had responsibity for preparing and publishing releases of Funnelback software, possibly including productised versions of new capabilities I might develop.

Brett Matson describes his arrival in the P@NOPTIC team:

> I joined CSIRO in Feb 2002. The situation as I recall was that my first role was in Kerry Taylor's team, but there was some shock from Mike Kearney (I think) when I turned up on my first day. They weren't expecting me. I was worried because I'd just quit a job I liked (a developer at the Australian Bureau of Statistics) to take the CSIRO role. The role CSIRO had in mind for me had been given to Walter Stein and they didn't have a role for me. I then spent a number of weeks sitting in an office with little to do, so I worked on my Java skills. Then one day you visited, offered me the role in the search team, and the very next day was when we went out to install P@NOPTIC at the University of Canberra.

Brett was initially on probation. During the probation period he voraciously learned the ins and outs of P@NOPTIC and, as noted above, was useful from day two. Despite getting up to speed faster than I could believe, at his end-of-probation review, he criticised himself for how long it took him to learn the product. ☺

Brett became a very valuable member of the team, and was a founding member of Funnelback Pty Ltd when it spun off in 2005. A short time later he became Products & Services Manager. In 2007 he became Funnelback CEO and remained in that role until he left in 2020.

Brett Matson hanging out with Nicko McBrain, drummer in *Iron Maiden*. *Photo: Stuart Beil*

In Stuart Beil's farewell speech, Brett was one of the three people he singled out for [favourable] mention:

> We have been a great team. It is a relationship that has evolved over the years and is one that is built on trust and confidence in each other's abilities. Our skills have complemented each other and this has allowed Funnelback to grow and prosper.
>
> The key thing I remember about you as a young software engineer was your overwhelmingly positive, can-do attitude to get the job done. It is this attitude that has put you in good stead.
>
> After the spinoff, for many years we were the A team. We closed so many deals back then. Me in sales, strategically positioning a deal and you in pre-sales providing technical solutions and advice. Indeed, at one point we would do sales meetings via mental telepathy. You knew when I was going to defer to you and you were waiting with a response that just nailed it. Our sales travels hold some of the fondest memories I have.
>
> Over the years we have been good sounding boards for each other. It can be lonely at the top and to have someone to share thoughts, plans, strategies, responses, reactions, analysis and advice on all matter of things has been invaluable. We have achieved so much together. I thank you for the support you have given to me in the various roles you have filled over the years. The mantle is now yours.

Stuart resisting headwinds at the summit of Arthur's Seat in Edinburgh.

Stuart Beil was an economist and policy advisor with the Australian government. He represented the Australian government at the United Nations Climate Change negotiations in Kyoto and Buenos Aires before joining the Sydney Futures Exchange as Manager of Electricity Futures Markets. After a period of consulting on emissions trading markets and being involved with a number of start-up companies, he joined CSIRO as General Manager Commercialisation. As noted before, he was attracted to P@NOPTIC at the 2002 Pipeline Review and was largely responsible for the spinoff in 2005. He stayed with Funnelback until 2013, with several years as Chairman and Executive Director of Funnelback Australia and Chairman of Funnelback UK. Stuart established a very impressive record in closing deals for first P@NOPTIC, then Funnelback. His extensive list of contacts found me accompanying him to sales meetings at Lloyds Bank in London, the Australian Stock Exchange, and many other well known organisations.

In the early days, Stuart attended an IIR search conference in Sydney (with Andrew Templer).

> I had a good number of Funnelback brochures in my briefcase and during the conference lunch break I thought that the audience should also be made aware of Funnelback, so I decided to put a brochure on every participant's chair. Needless to say the FAST attendees were not happy and complained to the conference organisers. My presence at the conference was irritating enough for FAST, and this was the last straw. They tried to collect as many of the brochures as they could and asked me to leave. I refused. I said I had paid the attendance fee and was entitled to stay.

In his farewell speech, Stuart Beil recounts some of his business background

I left the Sydney Futures Exchange to set up a company with other partners to try to commercialise carbon trading, but ended up working on other opportunities in other sectors, doing corporate restructuring, equity raising and performing Board and executive management roles on a number of portfolio companies. This experience was where I really cut my teeth in business. I witnessed the best and the worst the business world had to offer. I saw a million business plans for a million new ideas being promoted by a million personalities, including some shady characters. I felt the harsh winds of commercial reality. I rode the emotional roller coaster of each setback or success.

...

I was looking for deals, for technologies to commercialise, and I approached CSIRO. I subsequently was hired by CSIRO as General Manager, Commercialisation. In that role I oversaw the systematic evaluation of hundreds of CSIRO technologies that were ready to be commercialised. Having seen so many business plans in my prior life, the evaluation exercise was quite harsh and some CSIRO scientists did not always like my frank feedback. At CSIRO I enjoyed complex deal structuring, raising of venture capital, putting teams together, spinning off companies and licensing out technologies.

In 2003, David Hawking presented the P@NOPTIC search engine to my team. The first thing that drew my attention was, contrary to popular belief, not Dave's professorial good looks, but rather the fact that his team was generating some revenue. Some would say that generating the first dollar in a new business is the hardest thing to achieve. It is a validation that there is a market need. I liked the business and it ticked many of the boxes I look for in a start-up company. It was in the dot com, online, digital, information space which was a growing market. There were few competitors and the team were a great bunch of blokes who were smart and who had a great attitude. The key building blocks of a sound business were there. The key was to turn it into reality.

I began spending more time with the P@NOPTIC team and helped them sell and put some financial

rigor around their operations. We were pioneers in hosted (cloud) search which is a great business model, even today. We secured long term blue chip clients back then like ASX, Westpac and AGIMO. We started making too much money and in 2005, CSIRO took the view that the taxpayer should no longer subsidise a commercially viable entity, so the decision was taken to spin P@NOPTIC off as a separate company. The name was changed to Funnelback and Francis Crimmins, Brett Matson and I took the leap of faith, with David Hawking following soon after.

I was living in Sydney at the time and each Monday morning at 5am I would drive to Canberra and spend the first part of the week with the team before driving back on Wednesday or Thursday night. The extra pressures of running a new company meant I worked even longer hours. I had to move back to Canberra in order to cope with the new role. ...

George Ferizis came to the P@NOPTIC team in early 2005, having just finished work on a UNSW PhD on FPGAs (Field-Programmable Gate Arrays). Although based with us in Canberra, he used to spend his weekends in Sydney, driving up on Friday evenings and down on Sunday evening or Monday morning. He told us of his shock at noticing, while overtaking a car, that the other driver was reading as he drove, with a book propped open on the steering wheel. George says, "The Friday evening/Sunday night drives were especially 'fun', it's probably a big contributor to why I try to drive as little as possible now!" An important technical challenge which George addressed was the problem of replicating servers. He wrote a fast copy utility which dramatically reduced the time to copy crawls and indexes from one server to another.

George also reminded me of a bug which plagued us, particularly on the AGIMO govsearch project. When searching on govsearch there would be one page which seemed to find its way into the top results regardless of the query. The unfairly favoured page would sometimes change when govsearch was recrawled but there was always a "lucky" page to annoy and embarrass us. It was so mysterious and so embarrassing that a prize was offered for solving the problem. George claims that he still keeps the bottle of wine as a memento. (The explanation was that a bug in resolving URLs meant that bushels of random anchor text would be erroneously associated with a single page.)

Says George:

> I also remember dressing very inappropriately for my first Canberra winter and the subsequent experience of eating nachos outside at the Purple Pickle. There was also improv theatre that Matt Sheppard had introduced me to, which I had taken up again in Sydney before moving overseas for a while. He formed a bit of a cluster there, and I believe at one stage Tom Rowlands also attended.

Matt Sheppard gained early exposure to P@NOPTIC while studying for a Bachelor of Software Engineering degree at ANU. A final group project used P@NOPTIC to index email from a server using the IMAP protocol. After graduation he worked with Peter Bailey at Synop until 2005 when he joined the P@NOPTIC team as a software engineer. For the first year (2006) of the Funnelback spinoff he was seconded to the company, and soon after joined Funnelback as a Senior Software Engineer, leaving in Mar 2008, and returning in April 2010 as Manager, Research & Development. He remained in that role until his departure in February 2021.

At Synop, Matt worked in the reseller section and was responsible for many of Synop's P@NOPTIC installations.

> I recall having to patch that `evalperl` exploit Brett and Tom found in all our servers...and of course, we had our own somewhat hacked version of `xml.cgi` just to make it trickier. We

also had our own custom document-level-security thing in those days, specifically for Synop's content management system (which was called `Sytadel`) – As far as I know that was the first document-level-security implementation with P@NOPTIC.

2010: Matt Sheppard presenting at a user conference

Matt Sheppard recalls some memorable R&D projects

The rewrite of the search interface – Mostly driven by Nico Guillaumin's work and your desire to reduce the overhead of Perl. In the long run, I think it being well designed made a really big difference (thanks to Nico's hard work). I think we regretted several times not taking enough time to build good tests for faceted navigation, primarily because it was hard to test it well without being able to run the `padre` indexer for a real integration test (something Luke Butters eventually built some good infrastructure for). I recall pressing Nico at the time to find some way to have the results page rendering be able to be done either on the server side or on the browser side. We never found a good way to do it at the time, but it eventually came to pass with the whole single-page-application thing,[a] and then the server-side-rendering push on top of that.[b]

Getting to a model where content could be dynamically added to the search indexes (by indexing small groups of documents and merging groups together into larger sets over time) – Cliff Henderson had the first try under the name continuous-updating, but unfortunately it was never reliable, perhaps partly because the framework it was built on OSGi[c] – something we never got experienced enough to work well with. I think Will Parkinson tried to use it for some project at some point and the failure of that project got pretty much everyone to swap off using it. Luke Butters, who initially tried to get on top of the OSGi stuff, eventually rewrote it from the ground up and got it to a super solid state which ended up being very heavily used (I think William Hill was the highest volume customer. The secrets to making it stable, other than building on a platform we were more familiar with, were building it so it could be run completely stand-alone without a running web container (which made it easy to build tests for it), and designing it to handle being suddenly terminated at any time and recovering on the next restart. Luke had a battery of tests which would run it in a loop adding/updating/deleting documents and searching over the resulting index, and then had another thread killing the process at random times to flush out instability. Over time, he fixed a ton of bugs in both our code for it (concurrency is hard, who knew?), and contributed fixes for concurrency issues back to a bunch of open source libraries we used. For a long time I wanted to move everything over to run on top of that updatable index platform (which we called push), but somehow we never had the resources to put into it.

The project to rebuild the administration side of the product (with REST APIs and a nicer single-page-application based UI) – Nico drove the first part of this just before he moved to the US, which was focused on parts of the user interface used by customers after implementation was complete (e.g. search query analytics, best bets configurations etc) – it was called the *marketing dashboard*. By all accounts everyone who looked at it found it really useful, but I believe there was a big issue with actually giving customers access to it (especially in Squiz implementations). Expanding that out to the rest of the product's implementation functionality proved to be a very slow process – In part because we had to formalise a lot of "just do it with ssh and a text editor" implementation practices into APIs, and in part because each new aspect of the product seemed to need lots of overhauling once we opened the can of worms. ☺ By the time I left we were getting pretty close to replacing the last bits of Peter Thew's perl admin UI, which was close to the first bit of the product I worked on back 2005!

Nico's efforts to heavily automate the hosting environment – This made a massive difference to the stability of the hosting environment and meant that the very small hosting team could manage a very large number of servers in an automated way.

[a] https://en.wikipedia.org/wiki/Single-page_application
[b] https://angular.io/guide/universal
[c] https://en.wikipedia.org/wiki/OSGi

Dave: What were the achievements you are most proud of?
Matt: There were plenty of nice technical achievements, and problems solved. The story I tell in interviews about the tricky technical problem I solved was a class-loader memory leak where three different decisions conspired to make things fail in a particular way, and tracking it down led to some interesting bugs I raised against open source serialisation libraries we used.
Dave: Were there any things you tried to achieve but didn't succeed?
Matt: We spent a long time talking about rewriting the crawler to avoid crawling everything again every time (especially in the hope we'd end up with a cleaner and more extensible design as we did with the search interface). We never had enough capacity to invest in it though – the closest we got was rewriting the core HTTP processing (which included things like form interactions), which was a big step forward, especially for Cambridge University's use case, and the ability to debug what the crawler was doing during implementations.

The longest running failure was actually getting into enterprise search. We eventually had a nice-ish model for document-level access control, which I recall writing design options for back just after the spinoff, but we never got to a happy point in terms of actually talking to enterprise systems. We never had the investment required to build our own connectors for every enterprise system under the sun, and the various third party connector options we went through (persistent systems, entropysoft, manifoldCF) all failed in practice for one reason or another (entropysoft was the most successful until they were acquired by Salesforce). In the year or two leading up to my leaving there was a new round of negotiation on that front with the German company Raytion,[16] who we'd been talking to years earlier and who I think might be able to do the job, but are probably priced such that Funnelback's existing customer base would never go for it.
Dave: Any comments on the people you worked with?
Matt: Nico Guillaumin was always amazing, and I learned a lot by working with him. He had a knack for picking the right approaches for solving problems we encountered, and aways held everyone to a high standard.

Luke Butters' capacity to learn his way around complex systems and solve complicated problems through persistence always really impressed me.

I still often think about how you'd tackle problems as well, Dave – Just in the last few weeks I've been trying to dissuade people in my team from throwing caches at problems where they could instead actually make the underlying system faster, which was roughly my answer to everything when I started with you. ☺

I feel like I should say something about many other people: Ben Pottier who's spent practically his entire career with Funnelback and knows every subtlety of implementing it now; Dulitha

[16] https://www.raytion.com/connectors

Ranatunga who did a great job leading the new admin UI development and took over the team when I left; Brett Matson, who I think shielded us all from a *lot* of chaos above; Shaw Xiao, Michael Carter, and Alwyn Davis, who I'm now working alongside at Instaclustr. I think we really missed a trick not getting Alwyn working on the Funnelback product rather than implementation stuff back in the day. And lots more who I'm sure I'd have a thousand nice things to say about if I took a little while to remember all their hard work. ☺

You might recall that Nico Guillaumin was supposed to go onsite to the PM&C department for some enterprise search implementation on the day Kevin Rudd was pushed out. They wouldn't let Nico (and I think Pete Levan) in, because they were worried about journalists trying to sneak into the department!

2005: The P@NOPTIC troubleshooting team: Tom Rowlands and Brett Matson. Some of the servers seen behind are running P@NOPTIC but the majority belong to CSIRO's grid computing projet.

Tom Rowlands was among the first cohort of ANU's Bachelor of Software Engineering (BSE) (1999 – 2002). Prior to that he had been heavily involved in a solar car project at Lake Tuggeranong College. He joined the P@NOPTIC team in November 2004 and showed a strong willingness to pitch in and be useful. He liked getting his hands "dirty" with hardware and cabling, but could also present a highly professional image when interacting with customers – see the photo above. Tom worked on maintaining the Solaris port of P@NOPTIC, on a metadata survey for AGIMO, and on providing P@NOPTIC with the capability to exploit user click data. He visited the Federal Court to support a P@NOPTIC project there.

He remembers that the first P@NOPTIC User Group meeting, in 2005, was really engaging and was impressed that P@NOPTIC users were willing to travel across the country to attend. Several of them were using P@NOPTIC in ways which the P@NOPTIC team hadn't thought of. He particularly remembers being inspired by Darryl Lewis's presentation on the cool things that the ABC were doing with P@NOPTIC, and Darryl's suggestion that queries in referrer URLs in web logs could be used to annotate target documents.

Tom also remembers George Ferizis finding an obscure bug in the string replacement library we had written to reduce the large memory requirements of standard Java strings.

Tom wasn't seconded to Funnelback, but instead stayed on in CSIRO, completing a part-time PhD on *Information retrieval through textual annotations* in 2011. His PhD work is described in Section 4.13.1 on Page 78. Tom left CSIRO in 2015 to work in the Commonwealth public service.

2005: Peter Thew explains something to Francis Crimmins, who looks a little doubtful.

Peter Thew joined the P@NOPTIC team at the same time as Tom Rowlands. He had previously worked as a contractor with Aspect Computing which became KAZ Computer Services. Peter provided a lot of the system administration on our growing collection of P@NOPTIC servers. We were somewhat constrained in our ability to spend money on our infrastructure. Some of our vital servers were provided by deskside servers sitting in offices. Later we bought rack mount servers but due to lack of bolts and brackets we ended up with a stack of four of them on a bench, separated from each other with blocks of foam!

Peter was also a maintainer of the many perl scripts responsible for the P@NOPTIC services. Embarrassingly for my self-reputation as a C-coding expert, Peter remembers:

> ... spending a day trying to debug the C indexer (PADRE) when it would fail after 1 or 2 hours of processing. Turned out it was an array index – starting at one and not zero.

Like Matt Sheppard, in 2006, Peter was seconded to Funnelback Pty Ltd for the first 12 months of its existence. He continued at CSIRO until December 2014.

Peter Bailey draws lessons from the S@NITY/P@NOPTIC experience.

1. Working to solve problems for people helps open up additional new and interesting science and engineering problems.
2. We could have saved ourselves some effort and confusion by steering clear of using existing popular brandnames and non-alphabetic characters in the original branding!
3. Creating a simple but compelling product interface that people can use and experience even before they've "bought" it allows amazing science and engineering to shine. Without the science and engineering, there'd be nothing creating value for people to appreciate, but without the user experience it's just hard for people to grasp what the concept of "search relevance" truly means.

The early days were really exciting for me as we collectively pulled together a search engine on top of Dave's PADRE technology and applied it to real world problems, not just TREC test collections. It was an exciting time to be working on search!

Peter Bailey's career path has intersected at many points with that of PADRE, S@NITY, P@NOPTIC and the associated research. He worked with me on the development of PADRE on the Fujitsu AP1000 in the early 1990s, came back to ANU in the late 1990s and was instrumental in the launch of S@NITY at ANU and in the design and creation of web and enterprise test collections. After leaving ANU again in June 2000, shortly before the end of the ACSys CRC, he spent a period working for Nuix Pty Ltd, and then as a consultant for ETC. However, he soon re-entered the P@NOPTIC picture as a reseller, through Synop Pty Ltd. After the winding up of Synop in 2005, Peter joined my team in CSIRO. After the Funnelback spinoff occurred at the end of that year, I was unwilling to be a project leader of a group working on things other than search. Instead, Alex Zelinsky allowed me to work on my own research projects in the area of web and enterprise search. Peter and I reversed our roles and he became my boss. Later, at Microsoft, Peter and I both reported to Nick Craswell. ☺

2002: Peter Bailey at a P@NOPTIC lunch held at Mezzalira on London Circuit, Canberra City.

5.9 P@NOPTIC Resellers

As indicated previously, our goal as researchers was to sell P@NOPTIC as widely as possible, while minimising the amount of time we needed to spend on customer sales and support. Signing up resellers seemed to be a way of achieving those simultaneous objectives, and we had received approaches from two potential resellers, Synop and CiTR.

Synop Pty Ltd was founded by Nathan Wallace[17] and Peter Bailey served as its Canberra office manager for four years.

With Peter's experience with the P@NOPTIC team and its technology base, SYNOP was well placed to become a P@NOPTIC reseller. You can see from the green tablets on the timeline on Page 88 that they were quite successful in this role. Unfortunately, Synop wound up in late 2005.

Another reseller signed up at a similar time to Synop was CiTR, with a background in the University of Queensland. CiTR was represented at the time by John Gottschalk and John Scutt, and sold a product called AccessPoint. The relationship with CiTR was not as smooth as we had hoped and the way AccessPoint operated did not allow P@NOPTIC to show its capabilities.

[17] Nathan went on to become Global IT Manager for the Johnson & Johnson pharmaceutical company, and later started his own company Turbot.

5.10 Other P@NOPTIC Projects

5.10.1 Universities

P@NOPTIC did quite well for a while, signing up Australian universities. The University of Canberra[18] signed up at the same time as the University of Sydney. The University of Wollongong (Robert Robinson) chose P@NOPTIC early in 2003 but insisted that it had to run on a Solaris machine. Amazingly, Nick Craswell was able to port most of P@NOPTIC to run on Solaris and we went down to Wollongong to do the installation. ADFA signed up in 2003, and the University of New England (Gordon Smith) in 2004. We had a small search project with the Administration at the University of Queensland.

For a while we also provided a plagiarism detection service for the Australian Defence Force Academy (ADFA). One of their academics had determined that the major source of plagiarism at ADFA was copying of other ADFA assignments, either from the current year or from earlier years. They made a collection of submitted assignments and indexed it with P@NOPTIC. For each current assignment, P@NOPTIC extracted a ten word query, and that query was run against the index. Apparently that worked well until students started making more use of online resources.

Unfortunately, the release of the Google Search Appliance in 2002, later followed by the Google Mini, dramatically slowed down sales to Australian universities. Eventually LaTrobe, QUT and Macquarie University became customers, but we failed in multiple attempts to get business from Monash, RMIT University and the University of Melbourne.[19]

In Chapter 6 we will see that a focused approach on universities, after Google stopped selling the Google Mini, led to Funnelback signing up as many as 57 universities in the UK. Funnelback US (see Page 210) was also able to sign up many North American university customers.

5.10.2 Overseas Customers

2007: My second visit to Richard Rogers at the University of Staffordshire – our first UK customer.

During the cottage industry period we were approached by quite a number of overseas organisations, including quite a lot from Canada, the USA, and the UK. We were shortlisted in a competitive

[18] Katerina Christenson and Karl Maftoum were the UC people we liaised with.

[19] Brett Matson tells me that all of these eventually became customers after I left Funnelback. Never give up! Never burn bridges with prospects!

process to acquire a search engine for the Swiss national government in Bern. They were impressed but were frequently deterred by the lack of a local support office.

Some overseas prospects found us through the site `http://searchtools.com` operated by Avi Rappoport in San Francisco. Avi describes herself as a *search maven*.

2003 Infonortics Search Engines Meeting, San Francisco. From left: Geraldine Wade (Centers for Disease Control), Avi Rappoport (SearchTools), ?

The University of Staffordshire in the UK, contracted with a company to build their web presence. That company recommended P@NOPTIC for the search function, and lo and behold, we acquired the university as a customer in January 2004. I wanted them to feel supported and therefore, on a trip to the UK, arranged to drop in on Richard Rogers. After lengthy discussion about using P@NOPTIC to best advantage, Richard and a colleague took me to lunch at a charming, ancient pub out in the country, before I resumed my train journey to London.

Another very early UK customer was the Scottish Commission for the Regulation of Care (SCRC). They were responsible for registering and auditing the thousands of organisations providing care for children and the elderly in Scotland. I visited their Dundee headquarters more than once, and on one occasion performed an upgrade on their Windows installation, during which I was able to call Kat Ng back in Canberra, for technical support.

Eleanor Morton was the relevant manager at SCRC. Once when I rang to arrange a time to meet, she said that Saturday morning would suit them best. I caught a train from Glasgow. In the taxi on the way to Queen St station the driver asked me where I was going.

> **Dave:** To Dundee.
> **Driver:** That'd be for business then?
> **Dave:** That's right.
> **Driver:** Aye, Ah thought as much. Ah couldnae think o' any other reason fa going tae Dundee!

I thought it quite remarkable that three public servants would come into the office on a Saturday morning for a non-critical meeting. My amazement intensified when I realised that Eleanor lived in England and had driven up for the meeting. Her technical person had clearly had a wild Friday evening, and seemed quite "wired."

> **Eleanor:** "You're looking a bit rough this morning, Dylan."
> **Dylan:** Aye, that's right. But he's been in Glasgow. He'll have seen worse!

5.10.3 Government Agencies

A major reason why AGIMO was willing to pay a significant sum for whole-of-government search was because individual Commonwealth agencies were able to use a "slice" of the `govsearch` service to provide search of their web content, whether it was one site or a long list of sites – for example the health department wanted to include the separate web domains they maintained for their campaigns on alcohol and tobacco.

The Australian Broadcasting Corporation (ABC) is not a Commonwealth agency but is commonwealth owned. We were introduced to them as a potential client by Peter Bruza and Rob McArthur from the DSTC (Distributed Systems Technology Centre) CRC based in Brisbane. Based on Peter's thesis, DSTC had developed a technology called GuideBeam which helped users by suggesting query refinements. Peter and Rob had gained interest from ABC's New Media team and suggested that there might be a possibility to combine GuideBeam with P@NOPTIC at the ABC.

In late 2001, armed with a trial crawl and index of ABC content, I made contact with Rob Garnsey (head of ABC New Media) and his Publishing System Manager Darryl Lewis. Rob told me that I was the most recent of a long line of CSIRO researchers pitching their research to the ABC, but that I was the first to offer a product. He said that the ABC was open to purchasing leading edge products, but had no mandate to spend funds commissioning research.

The ABC liked P@NOPTIC but were concerned that CSIRO might change its research direction and abandon P@NOPTIC, and nervous because we had so few reference sites. They insisted on an arrangement where the P@NOPTIC source code was kept in escrow, to be passed on to them in the event that CSIRO no longer supported them.

The ABC were a very useful client. In the pre-iPhone era, Darryl Lewis was pioneering the delivery of ABC content, particularly news, to phones running the WAP protocol. For a time, he used P@NOPTIC as a publishing engine. Darryl also used P@NOPTIC as an e-commerce engine in the ABC shop, and reported a significant gain in sales following the change. This was very useful marketing material for us later.

2004: P@NOPTIC won an iAward at the Australian Information Industry awards dinner in Sydney. Rob Garnsey, Head of ABC New Media is at front with Richard Ang (NineMSN CTO), Trieu Hoang, and me at rear, clutching the solid stainless steel trophy. *Photographer unknown.*

5.10.4 Corporates

Westpac

In around 2002, I visited a person who had responsibility for the Westpac intranet – the internal web sites providing information and resources to staff. Backed up by Susan Feldman's IDC reports (among others) I argued the benefits and savings that they could achieve by providing effective search. He stopped me in mid flight and morosely told me that there was no opportunity for savings because the bank had already sacked thirteen of the fourteen staff who used to handle internal inquiries.

That unsuccessful pitch was an eye-opener to me and forced me to confront how savings from productivity improvements were sometimes achieved. My argument would have been that effective search coupled with the new intranet would allow remaining staff to be more effective, but it was clear that my pitch was going nowhere.

In 2004 however, we were contacted by Peter Kerry of another part of Westpac. He was looking to acquire a search capability for the bank's external web site and had put an intern onto the job of reviewing the available search technologies. The intern had come across P@NOPTIC and found that it was cheaper than the competitors and had some strong features in its favour.

Trieu Hoang and I visited Westpac, armed with a crawl and index of Westpac's web sites. I pitched P@NOPTIC's capabilities using live examples on our hosted search of their site. Peter and the intern were impressed and signed up for a hosted search service with a high update frequency – multiple updates on weekdays. One of our major competitors had told Peter that the bank couldn't expect to get a reasonable search solution for under half a million dollars. When Westpac went with us for a fraction of that price, the same competitor burned their bridges by ringing up and abusing Peter.

Westpac originally contemplated hosting P@NOPTIC on a server they leased in a third-party data centre. I was flabbergasted when I found out that the machine in question was powered by an Intel 486, a decade-old design, and that the lease was costing around $10k per year! In the end we provided search of their external web sites as a hosted service with updates three times each weekday.

Unfortunately, our all-encompassing search caused temporary embarrassment because it uncovered old content hidden on the site, still published, and still linked to the main page albeit via a long chain of links. When a customer searched for `interest rates` they were able to find an old page, offering higher-than-current deposit rates. Embarrassing but not P@NOPTIC's fault – the same ancient content would also have been accessible via global web search engines.

During our presentation to Peter Kerry I had mentioned our interest in providing search over internal content at the bank – intranet web sites, email, and other document repositories. They found the idea of indexing staff email an absolute no-no due to privacy concerns. This reluctance led to my idea of a Corporate Mail Manager (CMM) and ultimately to the Cnawen project proposal. See Section 4.14.2 on Page 81.

On the way out of that meeting, Peter showed me a long line of two-drawer filing cabinets, dozens and dozens of them. He told me that until a couple of years before, all of Westpac's important documents were filed in those cabinets. More recently, things had gone electronic.

The Westpac web site search project was a success and Westpac did eventually sign up for search of their intranet.

Other banks

Like government agencies, the major banks outsourced many aspects of their IT service provision. At least one bank told me that there was no point talking to them, I would need to talk to their outsourcer. After talking to one such outsourcer, I realised that they saw the case for improved search productivity solely through the lens of their ability to increase the value of their contract with the bank. Because our licence fees were relatively small, I could see that there was very little chance of success.

At a CeBit conference in Sydney in the early 2000s, Trieu Hoang and I ran a P@NOPTIC stand as part of CSIRO's exhibit. Trieu had the brilliant idea of setting up side-by-side comparison interfaces enabling instant comparison of web site search offered by an organisation with what it would be if provided by P@NOPTIC. We focused on banks. On our screen, you could select one of the four major banks, to be presented with the side-by-side live comparison for that bank. Clients could enter a query in a single search box and compare the results for accuracy, presentation, and speed.

Brett Matson talks about the Westpac intranet project.

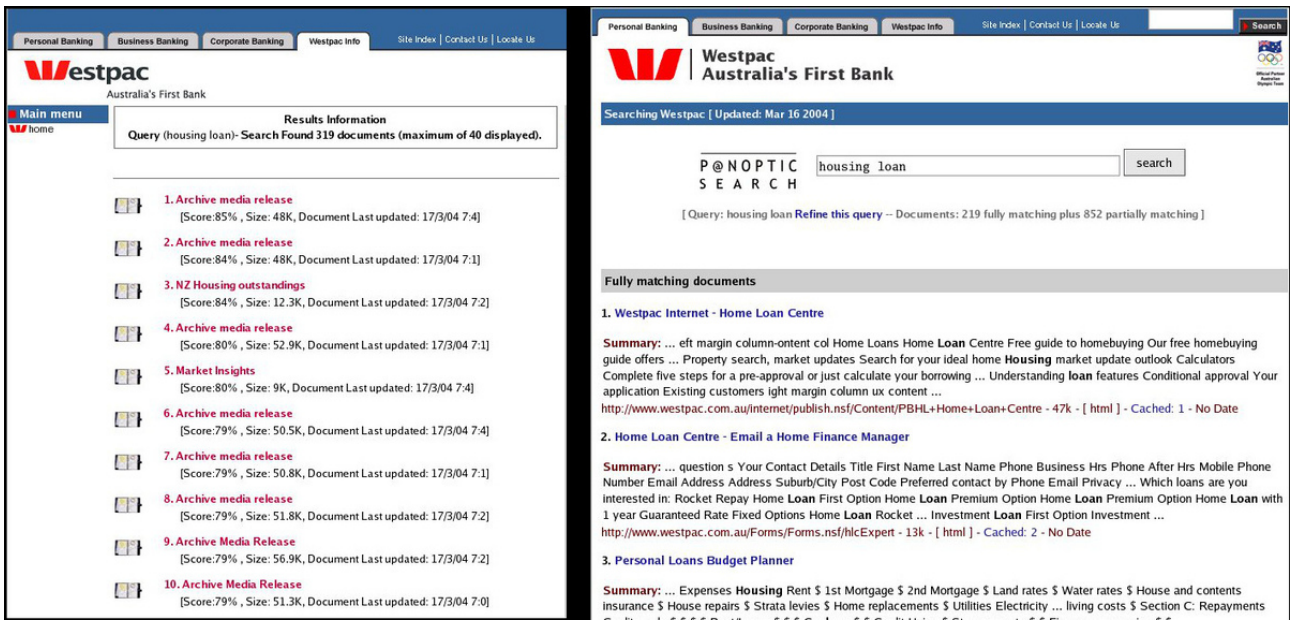
I installed Funnelback on their intranet around 2006 I think. I'm pretty sure the Intranet team won the prize for best organisational improvement for that project. Supporting it was a pain though. Two memories are:

- Their installation seemed chronically buggy and they were sending a lot of support tickets. After a while I found that even though we'd gone to a lot of trouble to build a working Windows installer for Funnelback, they insisted on installing it and then re-packaging it in their own installer, which introduced problems.
- Someone at Westpac decided that a search ranking algorithm that makes use of anchor text would be confusing for people, so they found the appropriate PADRE option and disabled it. Shortly afterwards they sent a strongly worded support ticket saying Funnelback's ranking is terrible. ☺

After a number of years, they decided they didn't have the budget to pay for the support on what was a perpetual license (about $2k per year), so they stopped paying it for a number of years. Several times they wanted upgrades/updates and we had to remind them that they'd have to pay for that. Eventually in 2016 they agreed to buy a new license again and have been using it since.

In 2011 they had a desire to do a large enterprise search project (mainly involving SharePoint), but didn't see it through.

Side-by-side comparisons became an important tool in both research and in sales. We produced a large number of different types of comparators, including ones with voting buttons or scoring pull-downs, and randomised placement of anonymised results. Side-by-side comparisons are relatively insensitive, but sensitive enough to reveal differences in ranking or interface which are large enough to influence a purchasing decision.



2004: The comparison between Westpac's incumbent search engine (left) and the P@NOPTIC alternative was dramatic for a business-critical query like `housing loan`. P@NOPTIC's third answer isn't useful but, thanks to the use of ranking features such as anchor text, the first two answers completely satisfy the intent behind the query.

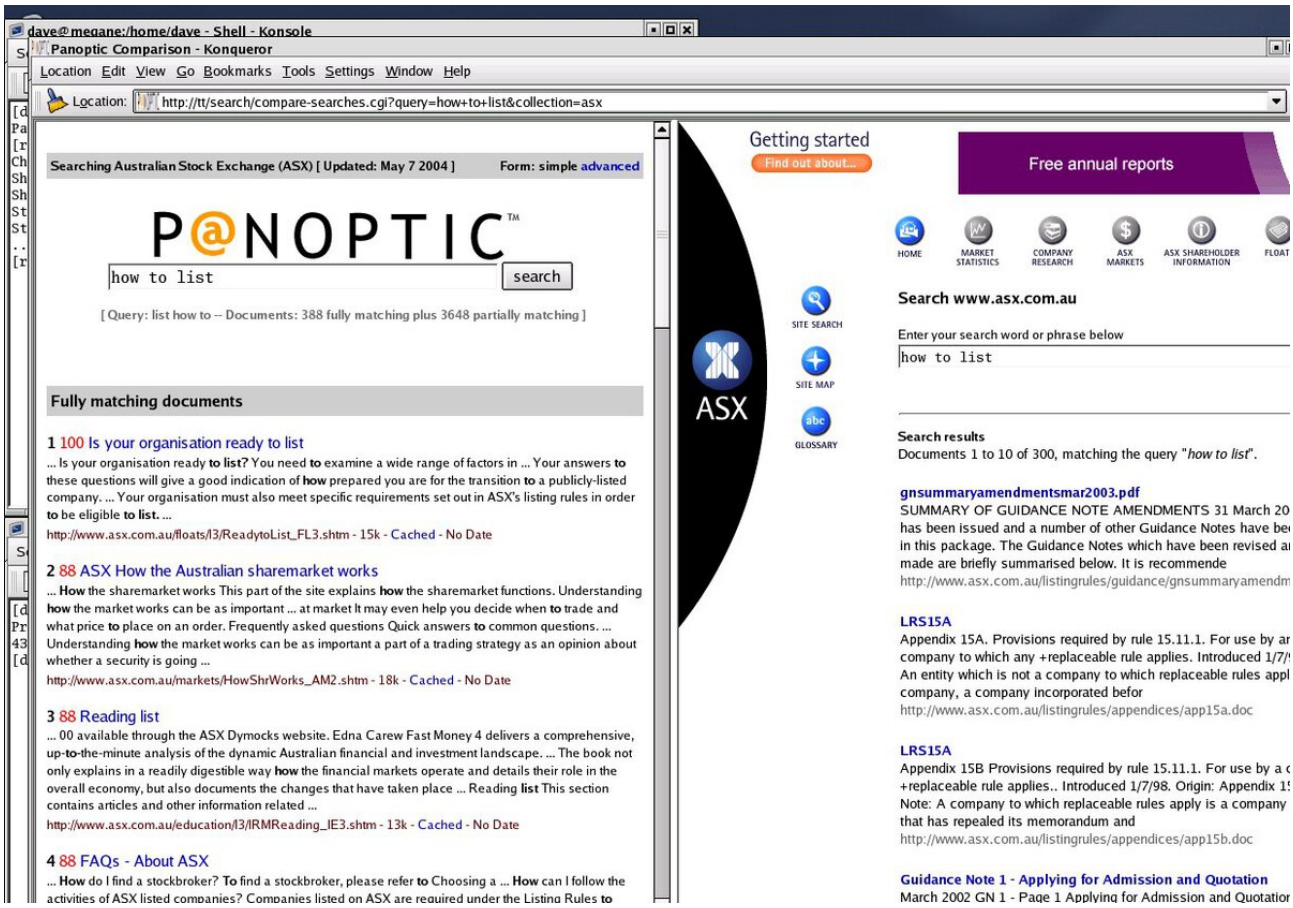
Australian Securities Exchange (ASX)

Stuart Beil established contact with Mary Ramsay at ASX via a friend with whom he had previously worked at the Sydney Futures Exchange. Stuart lined up a meeting with her and with the IT staff responsible for the ASX web site. We crawled ASX and set up a side-by-side comparator whose URL we gave to the attendees.

A few days later Tom Moschitz, Manager of Web Projects at ASX contacted us and said that they had fed about 100 queries into the comparator and found P@NOPTIC results superior on almost all of them. They would like to sign up for a hosted search service.

At the time the ground floor of the ASX building in Bridge St, Sydney featured a huge electronic stock indicator board and a roped off area for a coffee shop. The rest of the space held only a few chairs in which members of the public sat, in pathetic thrall to the changing numbers and the reds and the greens on the indicator board. The impression that the people there were quite naive was heightened when one of them approached us at a table in the coffee shop, wanting Stuart's advice (he was dressed in a dark suit and tie) on whether to take up the offer of shares in the demutualised NRMA. He, properly, told them that he couldn't help as he was not a licensed financial adviser.

Relative to government agencies, the ASX web site saw a much higher volume of search queries. Many visitors to the site wanted only to check the price of a particular stock, and in subsequent years we extended our query auto complete (QAC) system to include the current stock price whenever QAC suggested a company name or code.



2004: Side-by-side search comparison produced for ASX. The query is `how to list` and P@NOPTIC clearly meets the intent behind the query. In later side-by-side comparisons for other organisations, we took the trouble to emulate the organisations look and feel.

NineMSN

We were very pleased to sign up NineMSN as a customer in January 2004, since it was part owned by Microsoft! The task was to provide search over the various magazine sites like *Cosmo*, *Money*, and *Small Business*. Richard Ang was a director there and Anthony Wolf was our main contact. Their offices were in a very pleasant low-rise setting in Paddington. See the panel on Page 123 for more information about their installation.

5.10.5 Canada: The National Research Council

CSIRO's sister organisation in Canada is the National Research Council (NRC) / Conseil national de recherches Canada (CNRC). I was contacted by Estelle Vincent-Fleurs, the web manager at NRC, who had heard that CSIRO had developed a search engine which might help improve access to NRC information.

Travelling internationally for some other purpose in February 2003, I extended my trip in order to visit NRC in Ottawa to install P@NOPTIC on one of their servers, and to provide information and training. My arrival was something of a shock. The streets were covered in snow, and temperatures were below -30C at night.

I was looked after very well by NRC. One afternoon Glen Newton, who had expertise in open source and Java, promised to take me on my first visit to Quebec after we'd finished our discussions. We got talking, and it became quite late, but he drove me across the bridge to Gatineau, before immediately making a U-turn. I was delighted to be able to say I'd spent time in French speaking Canada – a total of 7 seconds!

We then went skating on the Rideau canal. Glen asked me if I had over-trousers, and when I said no, he said, "No problem, use these." It wasn't until we returned after skating around 10km that he complained of frost bite on his thighs and I realised that I was wearing his only pair. The skating was quite fun. Before winter, the city lowers the level of water in the canal, and wooden shacks hiring skates or selling beavertails (pancakes with cinnamon and lemon) and hot chocolate are towed onto the well-lit ice.

Estelle and Danielle Langlois took me to Tim Hortons for drive-through coffee, and wanted me to buy a Roots toque[20] – yes, they knew what root meant in Australia. Best of all was the Winterlude festival which featured stunning ice sculptures. (See the photo on the next page.) Next day I was taken to lunch with some senior NRC people at a restaurant near NRC. We walked, crunching across the snowy carpark in bright midday sunshine and a temperature of -27C. I was told that the sound of the crunch can be used to tell the temperature. I was also told that if I spat, it would be a pellet of ice that hit the ground, but I was too polite to test that claim.

I soon realised that, despite endless complaints, Canberra doesn't really have a winter.

5.11 Technology Developments

5.11.1 P@NOPTIC Licence keys

P@NOPTIC licences were restricted in various ways, such as by imposing a limit on the number of documents which could comprise a collection. I developed a simple licence key mechanism which encoded the parameters of the licence and attempted to prevent modification by the licensee. I wrote a tool for creating licence keys and implemented checks within PADRE. The checks were designed to minimize the risk of customer service problems. There was a lengthy grace period during which warnings were sent before service was cut off.

[20] A type of hat.

2003: Photos from my February 2003 visit to the National Research Council (NRC) in Ottawa. Top: Danielle Langlois and Estelle Vincent-Fleurs of NRC. Middle: The NRC audience for my talk. Bottom: An amazing life-size ice sculpture created for Ottawa's Winterlude festival.

5.11.2 Standard P@NOPTIC Installation

A standard P@NOPTIC installation included two views of each primary collection: Live and Offline. Each view stored a set of crawled data and indexes for that collection. Queries were served from the Live view and when the collection was updated, the new crawl went into the Offline view. Once crawling and indexing were completed, a series of licensee-configurable tests were run to determine whether the collection was of a plausibly correct size and that content was included from servers nominated as critical. If the tests passed, the Live and Offline views were switched. If problems were noticed post-switchover, the views could be switched back.

An improvement to the model was to deploy two servers, preferably in different locations. Updates would be performed on the secondary server and, when complete, the results would be copied to the primary. When the dual (or multiple) server model was combined with a load balancer, a failure or overload of the primary server would cause fail-over to the secondary.

5.11.3 *Efficiency

An advantage of the dual server model was that resource contention between the crawler, indexer and query processor was avoided. In the P@NOPTIC era, the crawler, the indexer and the query processor were highly dependent on availability of RAM, which was generally in short supply.

The crawler was written in Java and made use of Java objects and strings while leaving memory management to the Java runtime and garbage collector. The crawler needed to make a list of URLs already encountered, and the use of Java strings more than doubled the amount of memory required. As the crawl progressed, the amount of memory actually needed rose steadily and garbage collections took longer and occurred more frequently. Memory access patterns were essentially random.

The PADRE query processor also made essentially random accesses into the index data structures. It relied on the page cache supported by all major operating systems to keep resident in memory the parts of the index which were most likely to be needed. When a new index was loaded (cold start) query response was very slow – many seconds for a large index. This was because large numbers of disk accesses were needed to load the necessary structures into memory. We developed a warm-up script which could be used to make sure the important structures were loaded before switching views. Once warmed up, and with a good proportion of the index loaded, response from PADRE was reasonably good, but much slower than Google.

It came as a shock to me when Jeffrey Dean revealed at the WSDM 2009 conference, that Google indexes had been entirely memory-resident since 2001. After that, I felt much more justified in asking for very large memory configurations when we purchased or leased servers.

A significant fly in our query latency ointment was caused by other activities happening on the machine handling the queries. Updates for the current collection or for other co-hosted collections, queries being processed for other co-hosted collections, copying of indexes between machines, all these things could cause contention for disk accesses or cause critical index structures to be paged out of memory.

Add to this the fact that running a query caused a number of perl scripts to be run, and each of them in turn caused an expensive start-up of the perl interpreter. We had tried using `modperl` to keep the perl interpreter running, but encountered some problems. I enabled the PADRE query processor to run directly via CGI – i.e. incoming queries would come directly to PADRE from the web server (usually `apache`) rather than via perl. For relatively small collections like ASX, query latency reduced by two orders of magnitude. I thought that this was exciting, but my colleagues weren't enthusiastic because some features were still to be implemented and because customising the user interface would be much more cumbersome and risky. In any case, I was the only one who liked coding in C.

Because our query response latency was longer than it should have been, we set up a monitoring server to track it. Clients like AGIMO required us to report it. Unfortunately, we didn't monitor the reports closely enough, and sometimes missed egregiously high latencies. On one occasion I discovered that we had set up a latency monitoring system for AGIMO which, every hour, simultaneously

fired a dozen agency, state-gov and fed-gov queries at the same `govsearch` server. Such a volume of simultaneous queries was far higher than ever seen in practice and caused an wildly pessimistic view of our responsiveness. Response latency, particularly at the 95th percentile, became far more consistent and competitive when we were able to follow Google and keep indexes in memory.

Efficiency remained a major concern of mine until I left Funnelback in 2013. I was more concerned about it than any of my colleagues except for Stuart Stephen. I resisted any move to use virtual machines (VMs) in production.

Of course I lost that battle! Brett Matson says,

> On the virtualised vs. bare metal debate, I think there's a lot of merit in role based VMs (e.g. query processors, crawlers, admin, etc.). All the modern VM orchestration that enables auto-management, switch on/off, auto-scaling etc. would likely outweigh the benefits of running on bare metal. With `Engaging Things`, I can manage a sophisticated architecture with just a few clicks and I never have to worry about OS upgrades, updates, scaling, hardware failures or managing a hosting team. I've become a fan of modern cloud. ☺

While working at Microsoft I learned that Bing runs its services on physical rather than virtual machines.

5.11.4 *Efficient scoping: fscopes and gscopes

Prior to the initial launch of govsearch, I'd been very concerned that the mechanism we had for restricting search to say Victorian state government would be too slow. It added mandatory metadata constraints to the query behind the scenes, e.g. `+u:vic.gov.au`.

To speed things up, I added the indexer capability to set `fscope` flags on each document based on URL patterns. I was able to find a way to add up to ten flags to each document without increasing the size of the index. That was a convenient number because there were nine different jurisdictions. At query processing time the query form would insert a set of fscope bits to appropriately restrict the search. Any combination of flags was possible. For example, you could restrict answers to content from Western Australia, South Australia, Northern Territory and Federal sites.

Later, even more ways of slicing and dicing were needed and `gscopes` were introduced which allowed thousands of flags and the ability to form gscope expressions in reverse Polish notation! Brett Matson had asked for gscope expressions and I told him that I would implement the reverse Polish version and leave him to convert infix notation into reverse Polish – I don't think he ever did.

5.11.5 Security vulnerabilities

The original P@NOPTIC team were entirely focused on developing the best possible capabilities in the shortest possible time. In today's environment, it seems hard to imagine, but we failed to consider the possibility that our online services could be exploited for evil purposes by malicious people.

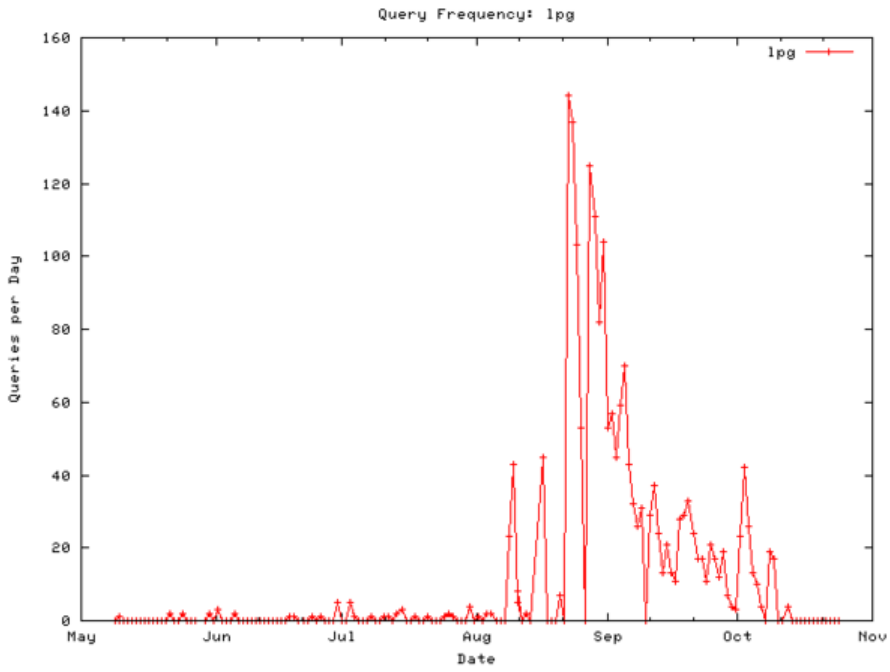
One risk pointed out to us by a hacker[21] at a government client was that of cross-site scripting (XSS). We were echoing the raw user query on the search results page. There was another gaping vulnerability:

Brett Matson recalls a serious security vulnerability

I remember sitting with Tom Rowlands one day when we were talking about `evalperl`, the then Funnelback search interface feature that let search masters write Perl scripts that executed within the search interface at query time. I don't recall how it dawned on us, but we eventually realised that it was possible to construct malformed queries that would execute arbitrary operating system commands as root! For fun, we made the search interface display the contents of `/etc/passwd`, and then promptly disabled the feature on all services.

Rest assured that this was in the very early days!

[21] Hacker in the most positive sense of the word.



2006: Submission frequencies for the query lpg on govsearch. In August 2006, the Howard Government announced a subsidy for LPG conversions of motor vehicles.

5.11.6 *Query spike detection

Francis Crimmins saw an opportunity to help searchmasters by building a mechanism for detecting sudden surges in popularity for a query or group of queries. The classic case occurred in 2006, when the Howard government announced a $1000 subsidy for conversion of vehicles to run on LPG. A flood of queries ensued for which there was no good answer on government web sites.

Fran's query spike detector relied on heuristics to determine whether a query had spiked and would email a list of designated alertees when a spike was detected. The idea was that an organisation could create or improve content in response to an unexpected surge in interest. A commercial business could potentially respond by expanding its stock or product range.

Alwyn Davis reminds me that Stuart Stephen added a more scientifically principled statistical test to the spike detector, based on his experience studying frog RNA! Alwyn was impressed by the cross-pollination of ideas from one field to another.

5.11.7 P@NOPTIC for Windows

Our reseller CiTR was keen to install P@NOPTIC on Windows rather than Linux platforms. As part of the reseller arrangements they undertook to produce a Windows version and supply it back to us. Unfortunately, no viable Windows version emerged. However, we had our own need for a Windows version, and Brett Matson set about creating one:

Brett Matson recalls the development of P@NOPTIC for Windows

It was nineMSN who insisted on the Windows port. Being part Microsoft, they didn't even want the perl front end to be visible to end-users, so they produced their own front end while we made the rest run on Windows.

The Windows port involved finding open source Windows ports of common Unix commands and having a config file that would select the appropriate executable based on the platform in use. I had trouble with these Windows commands because many of them had bugs since they weren't heavily used. I submitted at least one patch for one of the commands, but was soundly shot down by the community for not following correct patch-submission protocol. We also attempted to use Cygwin at

one stage, but it became cumbersome (i.e. in which situations should we use forward slash instead of backslash for path names?).

For the installer, I was reluctant to let go of all the work that had been done creating the RPMs for Linux and so fortunately found a straightforward way of extracting the files from the RPMs on Windows. One of the least enjoyable aspects of the Windows version was the need to integrate with the Windows IIS web server. Eventually it was all automated, but making system calls to change settings in a binary config file that I couldn't see always made me uneasy.

The Windows version became available in 2003 and went on to be used first by nineMSN and then by many other customers (e.g. Department of Infrastructure, Lifeline, and Australian Maritime Safety Authority). It was always questionable whether we should keep the Windows port, but last I heard it was still being supported in 2020.

5.12 Dave's Red Face Leads to Comprehensive PADRE Test Suite

In 2006, I was due to fly to Seattle to attend the SIGIR conference of which I was a Program Chair. Just before flying out I had completed a new version of PADRE which would bring major benefits. It passed basic sanity tests and successfully indexed and ran queries over our test indexes, so I threw the switch and pushed it live on our production services before dashing to the airport.

Bad judgment! Silly move!

By the time I landed in Seattle, reports had come in of serious problems with major customers. Behind-the-scenes query elements inserted by customer forms to achieve scoped search no longer worked. I skipped several conference sessions and sat under the cherry blossom trees in the quadrangle at the University of Washington, leaching off the UOW wireless network, debugging the problem and devising a fix. In an effort to avoid future recurrences, I wrote the first several of a now extensive suite of PADRE test scripts.

2006: Me in front of the cherry blossoms in the quadrangle at the University of Washington, Seattle WA.
Photo: Efthimis Efthimiadis

5.13 P@NOPTIC User Group Meeting

Late in 2005 we organised a first P@NOPTIC User Group meeting in the ground floor seminar room of the ANU CS&IT building. It was remarkably well attended, and among many others there were

inspiring presentations from the ABC on use of P@NOPTIC in news publishing and e-commerce, and from the Australian Antarctic Division on image search.

I had expected that I and other visible P@NOPTIC developers would be bombarded with questions about how to achieve certain things, how to address problems, and what was in our development roadmap. Instead, I sat and watched with increasing satisfaction, as users talked with each other about the technology we had created. Not only had we created a product which people used, but we had created a mutually supportive user community.

5.14 P@NOPTIC Server Fleet Prior to Spinoff

In a presentation I gave during spinoff discussions, I listed the complement of machines being deployed by the P@NOPTIC team. As you can see it was quite a large complement of somewhat ragtag machines. They were largely located in the secondary computer room on level 3 of the CS&IT building, but as Stuart Beil notes, some backup machines were located on the floor below:

> ...we were one of the first organisations to do hosted Software as a Service. We had redundant servers on the floor below connected by cables out the window. The ASX service went down one weekend and, with SLA penalties kicking in, we discovered that the CSIRO IT person was down at the coast!

Panoptic Servers currently (or soon) in use

Name	Purpose	Descrip.	Curr. Value
croc	exptl .au crawl	P: RM Coug Opt, 8GB, 2000 GB	$12k
crawlie	production crawls	P: RM Coug Opt, 4GB, 1000 GB	$7k
numbat	gov bureau	P: RM Coug P4, 4GB, 680GB	$2.5K
wallaby	agencies bureau	P: RM Coug P4, 4GB, 680GB	$2.5K
cuscus	general bureau	P: RM Coug P4, 2Gb, 680GB	$2K
dunnart	bureau backup	P: RM Coug P4, 2Gb, 680GB	$2K
old numbat		P: Dell Xeon, 2GB, 500GB	$5K
bilby	offsite failover	L: Dell Xeon, 3GB, 900GB	-
quoll	email alerts	P: Dell P4, 2GB, 500GB	$2K
gecko	mail-archive	?: Dell P4, 0.5GB, ?GB	$1K
bandicoot	demos	P: Dell Xeon, 1GB, 500GB	$1K
kangaroo	websites	L: Dell P4, 0.5GB, ?GB	-
platypus	release test m/c	L: Dell P4, 0.5GB, ?GB	-
wombat	backup	P: Dell P3, 0.5GB, ?GB	$0.1K
pleiades	Solaris testing	P: Sun, ??	$0.1K
thylacine	Research	P: EYO Ath, 1.5GB, 600GB	$0.5K
melbourne	Research	P: ACC Ath, 1.0GB, 600GB	$0.5K

2005: P@NOPTIC infrastructure at time of spinoff. L – Leased, P – Purchased, Coug – Cougar, a Canberra based PC manufacturer. `Thylacine` was used extensively for research by Nick Craswell, Trystan Upstill and me. `Melbourne` was acquired through a CSIRO-University of Melbourne grant to support collaborative research with Alistair Moffat.

Chapter 6

The Funnelback Spinoff

When Funnelback spun off from CSIRO, Stuart Beil, Francis Crimmins and Brett Matson elected to join the new company. They received options on a few percent of the shares in the new two-dollar company at a strike price of one cent. The options could not be exercised until CSIRO exited the new business. CSIRO solely owned and totally controlled the company. The new Funnelback board comprised Steve Kirkby (chair), Alex Zelinsky, Glenn Downey, and Rob Sale. Rob Sale left the board after about a year.

2006: After the first Funnelback Board Meeting, we repaired to the private dining room at Canberra's prestigious Courgette restaurant. Back row: Rob Sale, Steve Kirkby. Front row: Andrew Templer, Frank Liebeskind, Stuart Beil and Alex Zelinsky

The board soon commissioned Steve Liebeskind of Sydney Capital Partners to prepare a business plan, and appointed his brother Frank Liebeskind as interim CEO. Frank made Francis Crimmins R&D Manager, and Brett Matson Products and Services Manager. The board appointed Andrew Templer as Vice-President, International Operations, and on Andrew's advice, bought a licence for `RightNow` customer relationship management (CRM) software.

I wasn't happy with the spin-off arrangements and elected to stay in CSIRO but was seconded to the company about 50% of the time. I regretted that the proposed commercialisation model divorced

the product and the business from the Research – the end of my virtuous cycle. I wanted to continue my scientific research but I also wanted to work toward the success of the company of which I was the technical founder.

Alex Zelinsky, then Director of the CSIRO ICT Centre, told me that, although I was staying in CSIRO, he would make sure that I wasn't disadvantaged relative to the other founders. Eventually, he had to work very hard to make that come true but to his credit he did. He persuaded the CSIRO COMEX committee, the CSIRO CEO, and the CSIRO Board.

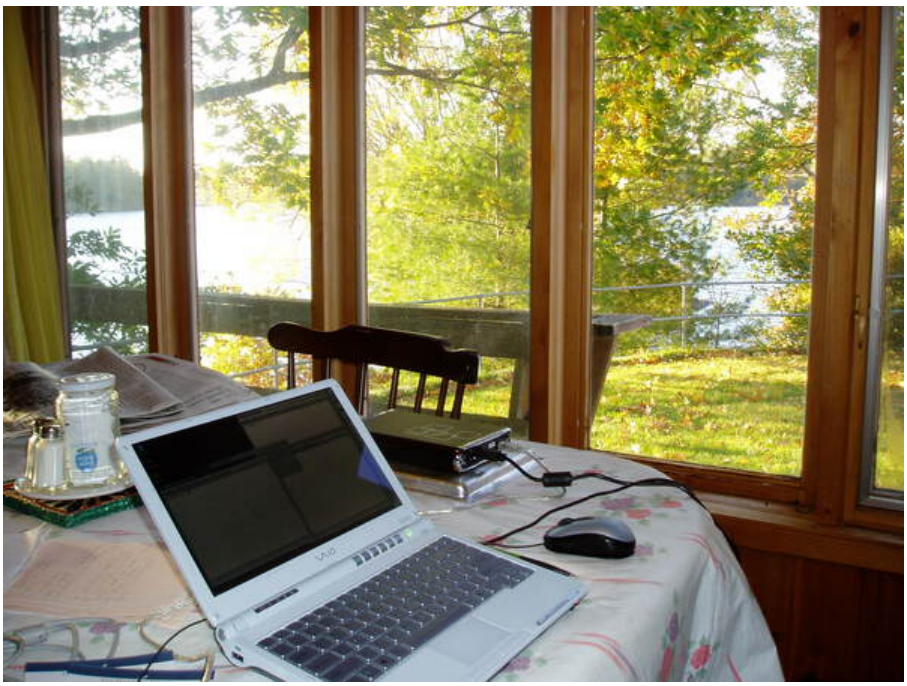
I also thought that founders should have been given a larger equity in the company, and I thought that CSIRO had given itself too much control. I couldn't help noticing that Sergei Brin and Larry Page still owned 34% of Google, even after large-scale external investment, and the IPO. I also noticed that if I had been able to continue in ANU, the P@NOPTIC inventors would have been entitled to 33% of the total return to ANU. In CSIRO, there was (temporarily) an inventors reward scheme providing a 0.5% reward, up to a lifetime maximum of $500k. My memory is that even that was overturned by the Education, Science, and Training Minister (Julie Bishop) who believed that work done by paid employees should not be further rewarded.

Members of the P@NOPTIC team were given the option of being seconded to Funnelback. Matt Sheppard and Peter Thew were seconded full-time for a year. I was seconded 50% of full-time and CSIRO told me that I should maintain separate laptops and email addresses. I confess I found the idea of two laptops unworkable.

2005: Working in a temporary outdoor office at the Trümmelbachfall near Lauterbrunnen, Switzerland, still with a clunky CSIRO-issued laptop. ***Phtotographer unknown.***

By that stage, I was using my own personal laptop (running Linux) for all aspects of life – I was a pioneer of BYOD (Bring Your Own Device). It was a Sony VAIO TX17 and it had a tiny but high-resolution 11-inch widescreen display [1] and weighed only 1.24 kg. With the optional larger battery, battery life was up to 14 hours. Remarkably, it had a good range of ports and a built-in DVD burner. In an office it was quite capable of driving a large external monitor. I called it my *Economy Class Laptop* because, unlike the clunky, heavy, CSIRO-issue, non-widescreen laptops, I could use it in an economy class aircraft seat without danger of it breaking when the seat in front crashed back. In those days I was doing a lot of international travel and used it to work on the plane for far longer than the standard-issue battery life.

[1] My eyesight was much better in that era than it is now!



2006: My trusty "Economy Class" Sony TX17 set up in my "office" in Thousand Islands, Canada. Maybe I shouldn't have been working while on holiday, but if you decide to work you might as well do it in pleasant surroundings.

George Ferizis declined to be seconded:

> No, I wasn't seconded to the spinoff. I distinctly remember having a conversation with Stuart Beil over coffee at the ANU about secondment where I elected to stay at CSIRO. I'm struggling to remember timelines after that, I know that I still worked on Panoptic/Funnelback but I can't remember what the arrangement was between the spinoff and CSIRO.

Instead, George worked on the *My Instant Expert* project which provided a question-and-answer service based on Wikipedia. You could ask a question from a mobile phone supporting WAP[2] and receive a concise answer.

Funnelback's offices were in the Epicorp incubator in the buildings originally occupied by CSIRO's Division of Computing Research.[3] We took possession in very early January 2006, and network connectivity was very soon set up. Once my Funnelback secondment commenced, I started working 2.5 days a week in Epicorp. I found that everyone else was still working in their CSIRO offices, and I spent a very productive time by myself.

6.1 *Multilingual Funnelback?

I used the time of peace in Epicorp to convert PADRE to use UTF-8 in order to permit us to address all linguistic markets, including mixed ones – ANU is an English-speaking campus but it publishes web pages in many languages and character sets. The conversion was a major exercise but well worth doing. PADRE needed to detect the character set used by a document and to convert it to UTF-8. The open source `iconv` library was a godsend and did all the conversion work. Eventually, we addressed the linguistic challenges posed by German alternative orthographies for accented letters, the different ways of writing Māori macrons, and the need for accent conflation. You needed to allow people to find accented content even if they didn't know how to produce accents on their keyboard. You needed to be able to match `salé` with `sale` even though that means a query for `poisson salé` (salted fish) also returns `poisson sale` (filthy fish).

PADRE eventually developed the ability to decompound German nouns such as:

[2] `https://en.wikipedia.org/wiki/Wireless_Application_Protocol`
[3] See Page 146.

```
bundesbankspräsident
```

(extending a method from Jacques Savoy to use bigrams), and to recognize and convert simplified and traditional Chinese. Wearing my CSIRO hat, Liyuan Zhou, Paul Thomas, and I conducted a comparison of alternative methods for segmenting Chinese.[4]

Becoming multi-lingual was a major challenge, potentially affecting all Funnelback components. My worst challenge was posed by a late 2006 opportunity to provide search for Israeli government web sites. Hebrew content is written right to left, but this causes no problem on the current web – characters appear in logical order in the HTML and it's up to the browser to display them right-to-left. Unfortunately, Israel is a technologically advanced nation and their government started publishing on the web before browsers reliably supported right-to-left display. They used so-called *visual Hebrew*. In this deplorable mode, the publisher works out how many words can fit on each line, and reverses the order of the characters in that line. For PADRE to index the content, it had to detect which lines were in visual mode, and reverse the order of the characters. But doing so was a nightmare because English (left to right) was mixed with Hebrew, and there were messy interactions with HTML constructs like tables and images. Before succeeding in this conversion project I went off to hospital for breast cancer surgery, and by the time I returned to work the Hebrew opportunity had passed.

2011: Feeling very grateful to Funnelback, I spent five weeks as a visiting professor at a university (IRIT) in Toulouse, France. I worked on Funnelback matters in the early morning and went to IRIT in time for lunch with colleagues. My host Mohand Boughanem attempts to perform an initiation ceremony at his house. ***Photo: Kathy Griffiths.***

6.2 Funnelback People

Michael Carter joined Funnelback from DSTO in mid 2006 and left in mid 2009 to work for the ACCC, just before the Squiz takeover.

> When I joined it wasn't a very large team. I remember Brett Matson, Francis Crimmins, yourself, Stuart Beil, and ringing in from the CSIRO, Matt Sheppard and Peter Thew. As time went on, some others I remember working with included George Bills, Ben Pottier, Kat Ng and Deniz Saticieli in marketing.

[4] https://david-hawking.net/pubs/cn-segmentation.pdf

Some pieces of work I recall are learning a lot of Perl to deal with Nickscript, re-writing the whole admin interface. Peter Thew complained about the offensive orange colour, and disliked "grinning idiots" (stock photos of people) in marketing materials and web sites.

I also worked on PADRE, getting add-on collections to work. I remember reading incredibly long C methods – one had about 35 different return statements in it!

I recall the original working space in the incubator building [Epicorp] which looked like it was about to fall apart at any moment, and the somewhat nicer office location in Dickson.

Funnelback was the first place I worked that had laptops as standard issue instead of desktops. That seemed pretty revolutionary at the time. ☺

Funnelback was a decent place to work. In retrospect, it could have done more to promote a positive working culture there. Not that it was bad, just that I've since seen better. I did enjoy the frequent lunch expeditions with most of the team. Even if we had to walk a long way to get anywhere on the ANU campus.

2006: Michael Carter and George Bills.

Ben Pottier joined Funnelback in June 2007 and has worked on Funnelback projects ever since. After completing his IT degree in Bruges, Belgium, he came to Australia, and

... found Funnelback through some obscure government bulletin board and I somehow convinced you all to sponsor me after 5 months of working on my holiday visa (FB definitely beats fruit picking!)

Ben describes his interview with Brett as follows:

Brett asked me three questions:

1. Do you know Linux? Kind of (I just said yes).
2. Do you know Perl? A little (I said yes).
3. Can you start Monday? Definitely!

It's been a hell of a ride since then!

Ben has worked as: Technical support engineer, consultant, senior consultant, search technology specialist, head of technology, solution owner/chapter lead, development manager. He's spent the past eleven years at Funnelback UK. (See Page 174.)

Ben Pottier remembers some interesting projects from his 14 years at Funnelback.

CareerOne with the first implementation of facets. My first breach of proper deployment procedures too. ☹ We did a Perl code deployment on some Australian morning and something went wrong. I immediately knew what part of the code had an issue and the fix, so I applied it in real-time and it worked. The client screamed at me over the phone to roll back and never to fix anything in live again during a deployment window. Processes are there for a reason after all!

AGOSP (Australian Government Online Services Portal). Working with Dave Doherty and Gordon Grace (him being on the AGIMO side) was a blast. A technically challenging build for us, but we were just a tiny component of this massive project. I remember going to project board meetings and it would have directors from AGIMO, from ATO and other major agencies.

GRDC (Grains Research and Development Corporation). Working on their site search, their team was just lovely and were great advocates. I still get their monthly newsletter though I frankly don't have much to do with the grain sector. Still mates with their team lead to this day.

The **Australian Taxation Office (ATO)** proof-of-concept for indexing some XML files. A multinational competitor had been there for 6 months and was struggling to make it work. I rocked up one day with FB installers on a thumb drive and we spent the whole morning getting the installers approved onto the ATO system since they didn't allow foreign USB devices. In the afternoon we got FB installed, collections set up and crawled, and by the end of the day we had built a functional search with facets which the competitor hadn't been able to in 6 months. I don't think we won that one in the end but I'm not sure why.

Stuart Beil's recollection of the unsuccessful ATO project

With the ATO, the demo Ben Pottier did was for their Reference Manager. Reference Manager is used internally by ATO staff and CSRs (customer service representatives) to answer questions about tax, best practices and how to deal with the ATO. Questions are received and the Reference Manager application helps the CSR to locate the appropriate script, which is a detailed set of questions and answers that enable the CSR to respond to the query.

We indexed their content and, in real time, their CSRs could do a search and get relevant answers they could use to meet the caller's needs. A multinational competitor fiercely controlled the ATO account. They had problems getting the Reference Manager search to work, so Gabrielle Davies from the ATO got us involved and Ben set up a proof of concept in no time. Gabrielle liked the solution. I had several meetings with her to get Funnelback into the ATO, but I suspect the competitor played a major role in making sure this didn't happen.

Alwyn Davis joined Funnelback in 2008, coming from a web design business in Rockhampton. He worked in the Products and Services team but left in 2009 for the UK. (See Page 172.) He returned to Canberra in 2010, serving as Head of Products and Services for a year before leaving Funnelback in 2012. Alwyn recalls:

> I really enjoyed working with the people at Funnelback and I still remember going to the Dumpling House on Fridays or getting a beer with everyone at the pub on the corner that was constantly closing and re-opening.
>
> I came into Funnelback with very little practical IT experience or skills, but I learnt a lot from the people there – I remember seeing how unflappable Gordon Grace could be and trying to emulate that, how Annie Pritchard managed to combine professionalism with being nice, how thorough and detailed Steve Barnes was in approaching projects. I've taken those experiences with me and still look back to them as guiding examples. I also remember how approachable you were and always happy to answer dumb questions I had!
>
> In 2011, there was just a handful of us in the office over Christmas, so Shaw, Tim, Prathima, Mandhakini and I had a Christmas lunch. I brought a ham, and Shaw brought chickens feet! ☹

2011: Christmas lunch for Funnelback Canberra. Clockwise from left: Shaw Xiao, Prathima Chandra, Mandhakini Iyer, Tim Jones. *Photo: Alwyn Davis*

2010: Alwyn Davis (far right) running a Funnelback training session.

2011: Funnelback Team in the Canberra Times Fun Run. Gordon Grace, Steve Barnes, Annie Pritchard, Alwyn Davis. I tried out for the team but was unable to complete the necessary star jumps in training.
Photo supplied by Alwyn Davis.

2006: Stuart Stephen in Funnelback's Dickson office.

Stuart Stephen had a remarkably interesting career. He started off as a dairy farmer in New Zealand, and became interested in automating certain processes in the dairy. Having built process controllers for himself, he generalised them and formed a company to sell them to farms and agribusinesses across New Zealand. After a few bad experiences on remote farms, he changed from hardware to software and moved to London as a developer of commercial software.

Seeking a change in 2004, he enrolled in a PhD at the Institute for Molecular Bioscience at the University of Queensland, graduating in 2008, when he moved to Canberra. Stuart is in the extremely rare category of people who have a doctorate but no undergraduate degree! Somewhere along the line I believe that he was a professional cyclist, and that he and his wife ran a coffee shop.

In Canberra, he spent a short time working with the Funnelback R&D team. Like me, he was

very keen to reduce Funnelback response latency. Because of his interest in bioscience, it was no surprise when he took up an opportunity to join CSIRO as a bioinformatician. He told us later that, with his computing background, he was able to dramatically speed up some of the bioinformatics computations in his new job.

2011: Annie Pritchard at a lunch in Dickson.

Annie Pritchard worked for Funnelback from July 2007 to May 2014, with a small amount of time off to pursue her interest in art. Her roles included account management, marketing and business development.

2011: Annie Pritchard in front of her work on display in an exhibition in Sydney.

Stuart Beil praised Annie's contributions in his farewell speech in late 2013:

> I remember first meeting Annie when I interviewed her for an office admin role in the Epicorp building in mid-2007. Annie was wanting to get back into the workforce after having kids and she started working part time. It became obvious pretty quickly that Annie wasn't content to do only what was required of her in her employment contract. She could see so many things that needed to be done at Funnelback and launched herself into them with passion and enthusiasm. She did this selflessly and without being asked. This exemplifies her caring, pro-active can-do attitude. She brought her design skills to the fore by publishing brochures, launching version 2 of our web site and organising events, including several User Conferences.
>
> Annie is detail and process driven. She helped put in place many of the systems and processes, like SalesForce, that allowed us to grow to the next level.
>
> Annie demonstrated how versatile and talented she was. She was a good all-rounder. It was clear, however, that Annie had a thirst for more knowledge. I was at full capacity managing clients as well as chasing new business, so I asked Annie to manage existing accounts. This was my gentle way of introducing her to the world of sales. Annie grew in confidence and gained experience in an account management capacity. The foundations were now in place to move her more fully into a sales role.
>
> Unfortunately Annie decided to leave Funnelback in 2010 to pursue a career as an artist. Thankfully for Funnelback she was soon convinced to return in a full time sales capacity and has put her artistic talents on hold for now. Since her return, Annie has been one of Funnelback's most consistent and high performing sales executives and I hope that I have given you the knowledge, confidence and encouragement to believe in yourself.

Annie Pritchard's memories of her time at Funnelback

I have very fond memories of working at Funnelback. I worked there over a period of 7 years, the longest I have ever been with the one company. I actually left at the end of 2010 to pursue some other options but I returned mid 2011. I always felt valued and respected working for Funnelback. Both Brett and Stuart were fantastic mentors. The skills and abilities they helped me develop have empowered me to continue reaching my professional goals throughout my career. Their leadership created a culture of teamwork, enthusiasm and humour. There were always plenty of laughs! I feel privileged to have been part of the Funnelback journey, from the early days at the original CSIRO Black Mountain site (and the deafening cicadas during the summer!) through to the takeover by Squiz. I also feel very lucky to have worked with so many great people along the way, including Professor David Hawking, such a brilliant yet humble evangelist for Information Retrieval!

6.3 AGIMO Projects

The initial govsearch, agency search, and media alerting projects for AGIMO are described in Section 5.7 on Page 100. After the spin-off, the contract with CSIRO was novated to Funnelback Pty Ltd. It was extended at least once, and then put to tender again. Brett Matson remembers not leaving the office until 6am after an all-nighter working on the successful Funnelback bid. He recalls the way Funnelback worked with AGIMO in the years after the spinoff:

> After spin-off we did what seemed like dozens of continuous rolling projects with AGIMO for the next 2-3 years. We'd come up with a list of project ideas and offer them up like a menu where we'd present the benefits, costs etc. They'd select most, but not all, and we'd get them done and come up with some more ideas. I think we finally lost AGIMO in 2013 when the Department of Finance's (AGIMO had wound up prior to that, so it was being managed by Finance) budget was cut, I think because they were supposed to focus more on policy and less on service provision. Peter Alexander had moved on by that point too, so we'd lost our strongest supporter.

Over the years the value of the contract declined. The media alerts service and the requirement to index state and territory government web content were dropped. However, Francis Crimmins developed a Compliance Auditor product[5], along the lines of Paul Thistlewaite's 1998 AWESOME

[5]Initially called WCAG auditor, in reference to an Australian Government standard

proposal (Page 35). It earned a one-off payment from AGIMO and I understand it has earned more than a million dollars from other clients over the years. It crawled web sites, and subjected them to a variety of automated compliance tests, such as: Valid HTML; Presence of Title; etc. It then presented an online score card which enabled web masters to navigate up and down in the site hierarchy and see compliance scores for different levels, plus to view scores on all the dimensions for individual pages. They could examine an individual page and see the individual causes of non-compliance.

The Compliance Auditor was rewritten a number of times over the years and in around 2013 was renamed Accessibility Auditor. Eventually it was rolled into the Search product.

6.4 *Faceted Search

Marti Hearst of the University of California at Berkeley pioneered the use of faceted navigation in her FLAMENCO project[6] in around 2001. Facets provide an easy-to-use category-by-category overview of a large set of search results, and allow a searcher to narrow the results by selecting values within categories. For example, in a search for `television` on an e-commerce site, the categories might be: **brand** – Sony, HiSense, Samsung, etc.; **resolution** – HD, 4K, 5K, 8K; **technology** – LED, OLED; **screen size**; **price range**. One shopper may narrow down by first clicking on 8K in the resolution category, then select OLED from technology, and finally pick a brand. Another may first select a price range, then choose a brand.

I sat next to Marti at the SIGIR 2000 conference dinner in Athens, Greece. She had been a program chair of the previous SIGIR and I watched the smoke coming out of her ears during a speech by one of the new program chairs. Glass of ouzo in hand, he described the major problems they experienced until they threw away the reviewing software they inherited from the previous conference and developed something decent!

Marti Hearst and Search USA

Between 2009 and 2011, Marti was seconded to the Obama administration to advise on effective search, and I believe was in charge of Search USA (`search.USA.gov`) which provided whole of US government search. The web provides little information about what she worked on in her time there but I've no doubt that she provided great advice on search interfaces. In 2009, she published a book on *Search User Interfaces*.[a]

In October 2012, Stuart Beil and I met with Ammie Farraj Feijoo who had by then taken charge of `search.USA.gov`. She explained that US government search had formerly been called `firstgov.gov` and had been powered by Inktomi, running on 80 servers.[b]

Since the Obama administration had decreed a strong preference for purchasing open source software, Search USA was now a composite of a free `.gov`-scoped search powered by Bing, and deep crawls of some extra government content by Lucene/SolR. Ammie wasn't willing or able to tell us how the two services were combined. It was quite clear that there was no opportunity for Funnelback.

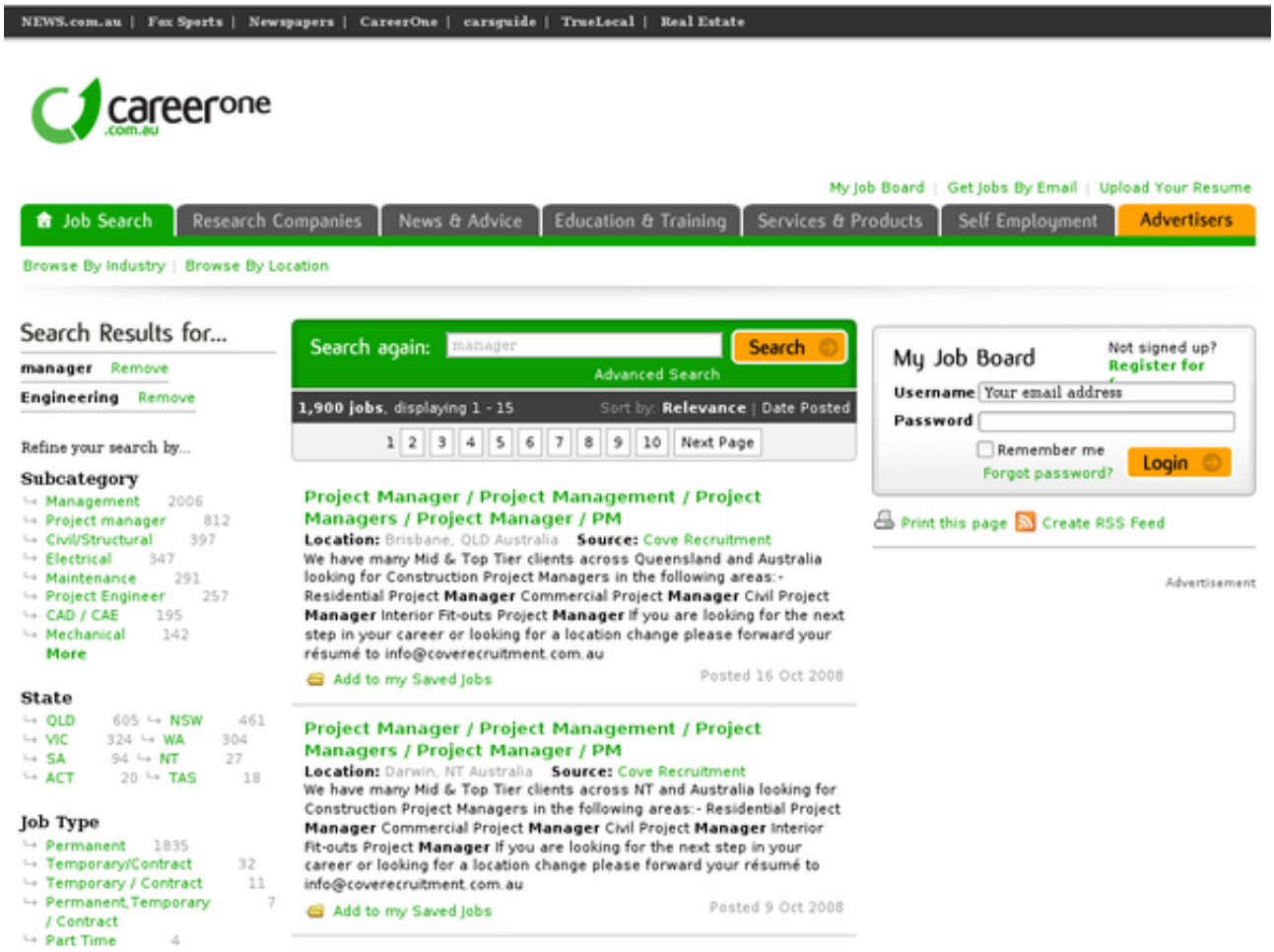
[a] `http://searchuserinterfaces.com/`

[b] Inktomi was founded by Eric Brewer, a friend of Marti's

For quite a while we provided hosted faceted search for CareerOne, an Australian recruitment site. They had bought a hugely expensive faceted search solution from a major Funnelback competitor, and paid two consultancy organisations to install it, but hadn't managed to make it work. We were able to set up a convincing service in a very short period. Brett Matson says:

> At the time that NetReturn Consulting brought us in to first meet CareerOne, Funnelback had no faceted nav capability, but within 8 weeks it was developed, released into the product, and the new CareerOne service had been implemented and deployed to production in our hosting environment.

[6] `https://flamenco.berkeley.edu/`



2008: An example of the Funnelback-powered faceted search interface on CareerOne.

Our CareerOne solution was responsive and effective, and ran very successfully for a period. Unfortunately, in 2008, they entered into a joint venture with US company `Monster.com` and transferred to Monster's platform. Brett again:

> After the Monster deal had gone through, Stuart and I attended a meeting with the CareerOne CIO, Simon Smith. He looked genuinely sad and said that the platform they'd be moving to was nowhere near the level of functionality or performance that we'd built for them.

Funnelback saw an opening for its faceted navigation capability in powering e-commerce sites. We gained some business from large retailers, such as Clive Peeters, who were just starting to see the benefit of Funnelback's technology, when unfortunately they went bankrupt. (Nothing to do with Funnelback I hasten to assure you, but due to an internal fraud.[7])

For me, e-commerce and job search sites provide the compelling case for faceted search, but many other organisations also see value in it. Brett Matson tells me that the Funnelback Higher Education offering[8] highlights this feature, and that it is in use at more than 50 universities.

Another major project reliant on faceting was done through Funnelback UK. The client was UBM Search Medica, a medical/health site based in the US. Their vertical search service covered many hundreds of sites around the world, as well as the entire content of PubMed.[9] Our index included more than 75 million pages – I believe that this was the largest ever Funnelback index. For some reason, in addition to providing facet counts, we had to return the results for the searches which would be triggered when a searcher clicked on each of the links. This meant running many subsidiary

[7] `https://en.wikipedia.org/wiki/Clive_Peeters`

[8] `https://showcase.funnelback.com/s/search.html?collection=higher-education-meta&profile=_default&query=courses`

[9] `https://pubmed.ncbi.nlm.nih.gov/`

queries in addition to the one the searcher had typed, and returning huge response packets across the Atlantic to the company's server in the USA.

To achieve reliable service and reasonable response latency in this challenging application, I persuaded Steve Morgan to buy two servers, each with 192GB of RAM – enough to allow the indexes to remain resident. The primary server responded to user queries and returned results. The secondary crawled the data and built the indexes, using the standard live and offline views. In the event that the primary server failed, the secondary automatically sprung into action in its place.

The service ran for almost 8 years, before the company was bought out.

6.4.1 *How Funnelback Faceting Works

The information needed to support faceted search must be associated with PADRE metadata fields. For example, **brand** metadata might be associated with metadata field `b`, and **price range** with `p`. The query `tv b:Sony p:1000-2000` could be used to retrieve the descriptions of all the Sony televisions in the price range. For faceting, PADRE needs to find the complete result set for the query,[10] examine the relevant metadata fields and return sorted counts for each different string for those fields. It's up to the search interface to provide all the functionality – displaying the categories and their values and counts, generating the PADRE queries corresponding to the set of category values clicked so far, and allowing the user to back out of earlier choices.

Returning accurate and consistent counts is not as easy as it sounds, but is very important. Users are affronted if they click into a category which should have five items and it only has four!

6.5 *Speling Correkshun

My former next door neighbour Tim Brook is irritated by the use of "spell correction" rather than "spelling correction". He's adamant that he doesn't need any assistance choosing the relative proportions of newts and toads in his potions.

As noted on Page 38, S@NITY included a rudimentary spelling suggestion mechanism based on `aspell`. It relied upon breaking the query into words and individually checking their spelling. I had difficulty persuading my R&D colleagues that this method was unsatisfactory but I saw it had the following limitations:

1. It misses better suggestions which are possible when the context of the whole query is taken into account.
2. It can't take into account the probability of the candidate suggestions. If two candidates are equally distant from what the searcher typed, it is better to suggest the one which is more frequently seen in the query stream.
3. It can suggest queries which have no answers in the collection.
4. It can't deal with errors which split or concatenate words.
5. It will suggest corrections for words which are correct in the context but not in the dictionary, such as local acronyms, people's names, or abbreviations.

   E.g. `ASIO dep sec` → `asian deep sex`

With an American dictionary the word-by-word correction system would not find even one of the more than 100 errors in the Pullet Surprise poem on Page 139.

[10] But only display the requested or default number of results.

Candidate for a Pullet Surprise

I have a spelling checker.
It came with my PC.
It plane lee marks four my revue
Miss steaks aye can knot sea.
Eye ran this poem threw it,
Your sure reel glad two no.
Its vary polished inn it's weigh.
My checker tolled me sew.

A checker is a bless sing,
It freeze yew lodes of thyme.
It helps me right awl stiles two reed,
And aides me when aye rime.
Each frays come posed up on my screen
Eye trussed too bee a joule.
The checker pours o'er every word
To cheque sum spelling rule.

Bee fore a veiling checkers
Hour spelling mite decline,
And if we're lacks oar have a laps,
We wood bee maid too wine.
Butt now bee cause my spelling
Is checked with such grate flare,
Their are know faults with in my cite,
Of nun eye am a wear.

Now spelling does knot phase me,
It does knot bring a tier.
My pay purrs awl due glad den
With wrapped words fare as hear.
To rite with care is quite a feet
Of witch won should bee proud,
And wee mussed dew the best wee can,
Sew flaws are knot aloud.

Sow ewe can sea why aye dew prays
Such soft wear four pea seas,
And why eye brake in two averse
Buy righting want too pleas.

Jerry Zar, 29 June 1992
Northern Illinois University
jhzar@niu.edu
Title suggested by Pamela Brown.
Based on opening lines suggested by Mark Eckman.
Published in the Journal of Irreproducible Results, January/February 1994

From `http://www.exodusbooks.com/Samples/Misc/SpellCheckPoem.pdf`

Before proposing an alternative mechanism, I developed an evaluation mechanism based on a testfile. Each line in the testfile contained an input query and a set of "correct" rewriting suggestions, or an indication that the input should not be corrected. Yes, the testfile included correctly spelled queries as well as erroneous ones. A scoring program reported the rate of good corrections, bad corrections, and missed opportunities for corrections.

My new method used a file of short strings, each with a probability of occurrence. The strings could be the same ones used in the query auto-completion system[11] – they included queries from recent query logs, and might include document titles, names from a staff directory, and anchor texts. Careful selection of the sources could avoid making suggestions which would result in "no results found." The latter is a major challenge when document level security is in force.

Very quick response was essential. We couldn't afford to calculate edit distances from an incoming query to every possible candidate in the strings file. Accordingly, we imposed a length difference threshold to restrict the number of candidates, then applied a rough filter to restrict the set even further. Finally, we calculated edit distances (an expensive calculation to do accurately) and combined a distance score with the prior probability of the suggestion. This process resulted in a ranked list of suggestions.

Configurable thresholds determined whether no action was taken, whether a "Did you mean?" was displayed, or whether suggested corrections were fed into the query blending process.

6.6 *Query Blending

At the WWW14 conference, held in Japan in 2005, Nick Craswell and I had lunch with Jan Pedersen, who was then Chief Scientist at Yahoo! Jan talked about the difficulty of "moving the needle" in search quality. In the early years, he said, search engines had made great strides in improving overall

[11] See Page 201

measures such as NDCG averaged over very large query sets. Recently, however, it had become difficult to improve overall average scores. You could identify a problem, invest heavily in a solution, make a small proportion of users deliriously happy, but see no visible change in the overall measure.

Jan said that *query blending*, in which variants of a user query are generated and run against the index, was one method that had brought benefits. In query blending, results from the variants are blended with those for the original query.

Query blending seemed much better than blindly replacing the original query with a "corrected one." I had seen this sort of behaviour before in results from Google – at one stage every single first page result for `Kathy Griffiths` was actually a result for `Kathy Griffin`[12] and every result for `David Hawking` was for `David Hawkins` or even `Stephen Hawking`. I thought that this aggressive query rewriting was an unfortunate over-reach. It helped people who couldn't spell but made it very difficult for people who had spelled correctly. In the worst case, it would make some web content unfindable and impose a "tyranny of the majority."

I implemented a Funnelback version of query blending, using one or more candidates from the new spelling suggestion system. To avoid the adverse consequences of falsely assuming that a query was misspelled, I designed a feature which meant that if there were full matches to the original query, the best of those would be guaranteed to appear at rank one, no matter how strong was the evidence in favour of "corrected" spellings.

6.7 *WARC files

Early on, Funnelback gatherers essentially replicated the directory structure of the data being gathered. For every document in the collection, there was a file on disk. When gathering from a database there was a need to insert artificial directories to avoid "too many files in a single directory" errors, or poor performance.

With big collections, there were serious problems with this file-per-document approach. A lot of space was wasted in the directory structure and operations over the gathered data, such as indexing or copying, were very slow.[13]

To support CSIRO and Funnelback research, the PADRE indexer was given the ability to index more compact and efficient representations such as tarfiles (tape archive files, no longer assuming tape) and various versions of WARCfiles (Web archives). PADRE could index compressed files whether in a directory structure or in a tarfile or WARCfile. WARCfiles were used when distributing the UK2005 and UK2006 data sets and, at my suggestion, the ClueWeb09 collection.

I did a review of the best way to speed up collection updates by bundling up files. I considered zipfiles, PDFs (quite like a zipfile internally), tarfiles and WARC files. I even made a preliminary design of IFFF, an Indexer-Friendly File Format, in which raw documents were filtered down to only the content which would actually be indexed. My recommendation was that WARC files were the best option. It was a published standard, specifically designed to cope with web content. It was easy to process sequentially, and you could read file headers to determine whether to process the content of a file or efficiently skip to the next header. If individual files were compressed rather than the whole archive, you could maintain pointers to each compressed file.

Acting on this recommendation, Cliff Henderson implemented support for the FunnelWARC format in the crawler.

6.8 *PADRE Index Structures

To achieve acceptably snappy response to user queries, the process of generating a document ranking by scoring documents in the index against the query (*ranking*) should take no more than about 100 milliseconds (ms), i.e. about 0.1 seconds. To that must be added the network latency, and the times

[12] Whoever she may be!

[13] The Google Anatomy paper talks about the use of *BigFiles* to avoid the slow down caused by disk seeks.

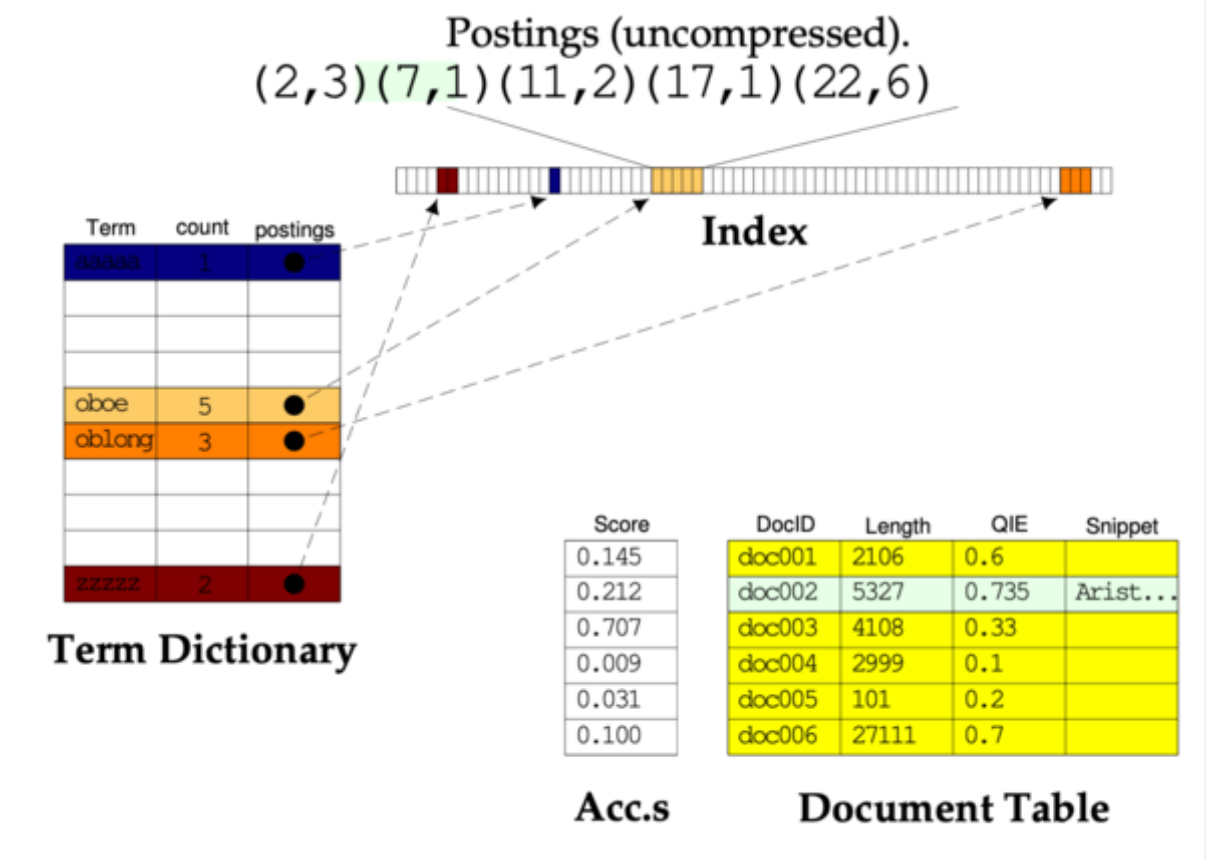
taken by the user interface layer, the spelling corrector, the snippet generator, and possibly other components of the system such as faceted navigation and related queries. If query blending is used, ranking must be done more than once.

The speed of query processing may be heavily influenced by the structure and content of PADRE indexes. The following panel shows a simplified version.

6.8.1 *Index compression

Compressing postings lists is a technique for speeding up query processing. It allows more postings lists to be kept in memory and speeds the process of reading them from disk. Initially PADRE used two schemes studied by Alistair Moffat: Golomb and Elias Gamma. Golomb achieved better compression but took an extra pass through the data. Accordingly we mostly used Elias Gamma.

What's a postings list?



Format of PADRE indexes. The Term Dictionary is an alphabetic list of all the words found in the collection. In this version, every document in the document table has an associated score accumulater. QIE stands for Query Independent Evidence score, e.g. PageRank.

Each word in the term dictionary has a postings list associated with it. Every occurrence of a word in the collection is recorded as a posting, which consists of the number of the document and the number of the word within the document. The fifth word in the third document would be represented as `(3,5)`. The order of the postings reflects the order of the word occurrences in the collection.
The example shows that if the query were `oboe`, documents 2, 7, 11, 17, and 22 would achieve scores. Information in the document table would allow them to be ranked and presented with query-biased snippets.

At SIGIR 2002 in Tampere, Finland, Hugh Williams presented a paper on vByte compression. It didn't compress as well as either of the other methods but Hugh showed that it was much faster to encode and decode because it used byte rather than bit operations.

I'm seldom able to sleep on aeroplanes and, on the flights back from Tampere, I coded up vByte

and confirmed Hugh's observations as we descended into Sydney. From then on, PADRE used vByte compression. However, in around 2013, Tim Jones and I supervised a Computer Science Honours project by Luke Butters which compared several alternative compressed index formats. Luke found improvements to both the form of the postings, and to the way they were compressed.

6.9 *TAAT

In common with most TREC systems, PADRE originally processed a query such as `the wizard of id` one *term at a time* (TAAT). This means that for each word (term) in turn, the index is used to find all the documents which contain that word, and update their scores. In the simplest model, each document in the collection is assigned a score accumulator, which is set to zero before each new query is processed. After all the query words are processed, the document accumulators are sorted into descending order, to give the ranking to be presented to the searcher.

Modern computers typically feature multiple processing cores. It is definitely possible to reduce the latency of ranking by using multiple threads but ranking is data intensive and it turns out that memory bandwidth limits the value of multi-threading. Furthermore, multi-threading may reduce latency when the query load is light but it doesn't improve maximum throughput.

The simple TAAT method works well for small collections, but for a large collection of, say, 50 million documents, it becomes impossibly slow:

1. Zero-ing 50 million accumulators takes of the order of 90 milliseconds on 2019 hardware. That uses up nearly all of the time budget for responding to a query![14]
2. All 50 million accumulators have to be examined to define the candidates for sorting.
3. The sort itself is very time consuming, even if scores are quantised and a counting sort method is used.

*Limiting the Number of Accumulators

An approach implemented by PADRE in the TAAT model was to limit the number of score accumulators, and provide a fast lookup mechanism (a hash table in PADRE's case) to find the accumulator associated with a particular document. Once all the available accumulators have been used, documents not yet assigned an accumulator cannot make it into the final ranking.

If you have only 500,000 accumulators for a collection of 50 million documents, then the time to zero them reduces by a factor of 100, and other steps are similarly speeded up. The downside of limiting accumulators is that careful management is needed to avoid a loss of quality due to good answer documents missing out on an accumulator. In the case of `the wizard of id` the score accumulators would fill up a fraction of the way through the set of documents containing `the`. Perhaps none of them contain any occurrences of either `wizard` or `id`.

The major technique used in PADRE was to order the query words by decreasing frequency of occurrence in the collection.

`the wizard of id` → `id wizard of the`

That technique made a big difference but there were still difficulties in deciding what to do with document fields such as referring anchor text, titles, URL words, ...

Overall, it was very hard to set the limited accumulator parameters to achieve low latency without detriment to the quality of results.

*Stop word removal

One efficiency approach used by ancient retrieval systems is removal of *stop words*, i.e. words which contribute little to the overall meaning. This technique helps deal with the limitations mentioned

[14] Andrew Trotman describes methods for speeding this up, but worst case performance is still possible.

in the previous section on limited document accumulators. When the query consists of hundreds of English words, removing words like `the` and `and` seldom makes a difference to the results but in short queries it can cause harm. Classic examples include, `to be or not to be`, `the THE operating system`, and `The The` (a post-punk band). The query `to be or not to be` was an early test of S@NITY. The original S@NITY took a long time (relatively) to process that query, but it did find all of the correct answers and, as far as I recall, no false positives.

In the early days, some P@NOPTIC customers demanded that we provided a stop word removal mechanism.

PADRE eventually did minimal stop word removal but only if the query contained at least a threshold number of non-stop words. We started with a combined English-French stop list, and allowed searchmasters to define their own. Eventually we added stop lists from a very useful multi-lingual resources site operated by Jacques Savoy at the Université de Neuchâtel, in Switzerland.

6.9.1 *Query Shortening

Ross Wilkinson was in Canberra soon after the launch of S@NITY. He demanded to see the search interface, and immediately copied three long paragraphs of text from a document that happened to be open, and pasted them into the S@NITY search box, chortling when S@NITY took an insane amount of time to respond.

After that I developed a query shortening mechanism for long queries. It made a list of the distinct words in the query, sorted by increasing frequency of occurrence, eliminated words with zero frequency, and retained only the k lowest frequency (most information bearing) words, where k was a configurable parameter.

6.9.2 *Stemming

A traditional information retrieval technique reduces query words to their stems. For example `goals` and `goal` would stem to `goal`. Depending upon how aggressive the stemming, `goaling` and `goaled` might also stem to `goal`.

Because some clients demanded it, I implemented the stemmer due to Martin Porter. However, unlike traditional IR systems, PADRE indexed words rather than stems, and added an efficient mechanism to map a stem to the relevant set of words. Indexing stems rather than words discards useful information, possibly preventing the discrimination of different meanings, e.g. `economics` and `economical`. A PADRE option allowed you to choose whether query words were automatically stemmed or not. If you chose not to automatically stem, you could use the hash operator to apply stemming to particular query words, e.g. `goal#`. Of course you could also use the right truncation (wildcard) operator `goal*` to match the same or a similar set of words.[15]

In the context of the very short queries submitted to enterprise web sites (on average less than two words), stemming generally harms result quality. An exception is plural/singular conflation – searchers don't know whether to search for `camera` or `cameras`. Eventually, PADRE defaulted to using a simple home-grown stemmer which provided only singular/plural conflation for English and French (with an eye on NRC, Canada). A search master could choose a heavier stemmer if they wished. Eventually, I implemented most, if not all, of the stemmers for different languages from Jacques Savoy's multi lingual resources site at the Université de Neuchâtel.

6.10 *DAAT

From Andrei Broder, I learned that Web search engines generally used a *document at a time* (DAAT) approach rather than TAAT. In this approach, the postings lists for all the terms are scanned in parallel. In the basic form of DAAT, you look for documents containing all the query terms (implied

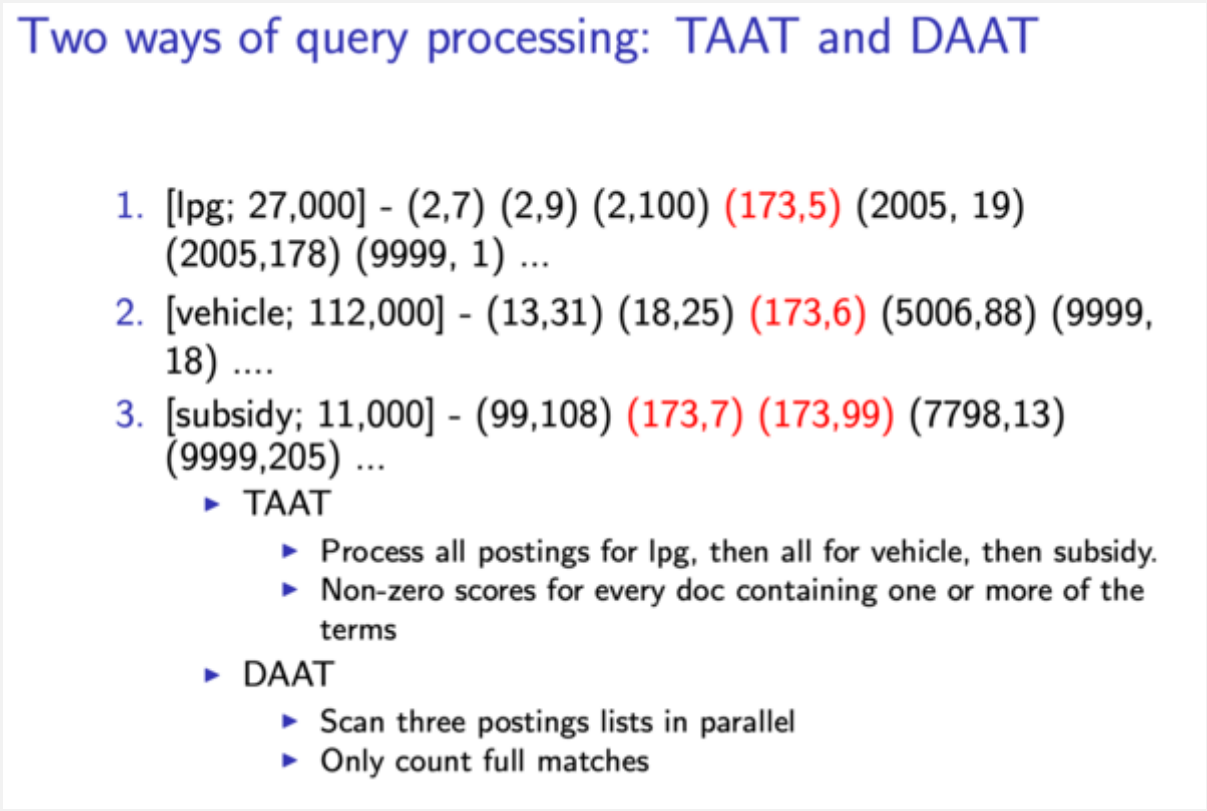
[15] Left truncation was also provided but was very expensive and seldom used.

AND). This means that you can quickly skip over sections of postings lists where an AND match can't occur.

When an AND match is found, you can compute its final score and insert it in a *heap*. Unlike TAAT, if processing is hit by a time-out, DAAT normally has a ranking ready to present.

I implemented DAAT in PADRE with a looseness parameter k. Instead of looking only for full matches, PADRE looked for matches with no more than k words missing. If you wanted AND you could set $k = 0$. PADRE kept separate lists of full and partial matches.

DAAT v. TAAT



A slide I used in one of my many talks, explaining the difference between DAAT and TAAT

6.10.1 *Speeding up DAAT: Skip blocks

Following another idea from Alistair Moffat and Justin Zobel, I implemented skip blocks in postings lists longer than a threshold length. Each block of say 1000 postings is preceded by a header which records the highest document number represented in the block and a pointer to the beginning of the next block.

This combines well with DAAT because, when scanning forward to find a particular document number, you can instantly identify blocks which can't contain that number and skip them. In a very long postings list you can potentially avoid decompressing millions of postings.

6.10.2 *Speeding up DAAT: Index re-ordering

As previously discussed, web search engines (and P@NOPTIC / Funnelback) rank using a combination of static and dynamic scores, where the static scores are derived from query independent evidence such as URL length, URL depth, click propensity and inlink counts or PageRank. Long and Suel[16] had the insight that assigning document numbers in order of decreasing static score (DSS order) could make DAAT scoring even faster. If you terminate processing early you know that all the

[16] https://www.sciencedirect.com/science/article/pii/B9780127224428500203

unexamined documents have a lower static score than any you have previously examined. Assuming that the static score is a reasonable percentage of the final score, it is quite likely that you have already seen the highest scoring documents.

I developed an index re-ordering tool and Tim Jones and I wrote a paper[17] in which we compared four different orders for several different collections. The orders were crawl-order, reverse-crawl-order, random, and descending static score order. We showed that full search quality could be achieved by the DSS index with far earlier cutoffs than for the others.

6.11 ✳Green Search

In Section 5.11.3 on Page 121 I have described how concerned I was with reducing latency in query processing by use of efficient algorithms and languages. I also mentioned my resistance to using virtual machines for production search services.

I had many debates with Francis Crimmins, who argued that it was human efficiency which was important. His claim was that programmers were a limited resource and that using high-level languages and programming tools to reduce development time was what we should be optimising. To counter this, I pointed to the high cost of running servers in data centers at TransACT in Canberra, and in the USA. That cost was approximately proportional to the number of servers. If we increased efficiency substantially, and reduced the number of servers, we could save at least the cost of hiring a programmer. If our scale of operations increased by a factor of ten or a hundred, as we both dreamed, the benefits of efficiency gains would be multiplied by the same factor.

I was also concerned by the greenhouse gas emissions generated by our servers. I took measurements of our servers, and found that their power consumption changed little between when they were idle and when they were working hard. I first presented my observations at the 2010 Squiz user conference in Wellington, NZ.

My presentation[18] presented various ideas for reducing greenhouse emissions. It proposed a multi-tenant Funnelback Search Cloud as opposed to dedicating virtual machines to individual Funnelback customers:

Emissions calculations for a representative server.

- Power consumption: 220 Watts
- Allowance for aircond: 73 Watts
- Total power draw: 293 Watts
- Hours per kiloWatt-hour: 3.41
- Energy used per year: 2567 kiloWatt-hours
- $C0_2$ emissions, Aust: 2.445 tonnes/yr
- $C0_2$ emissions, UK: 1.400 tonnes/yr
- $C0_2$ emissions, France: 0.225 tonnes/yr

Emissions calculations for a representative Funnelback server. ***From my presentation at the 2010 Squiz user conference in Wellington, NZ***

[17] https://david-hawking.net/pubs/HawkingJ2012.pdf
[18] https://david-hawking.net/presentations/Hawking_SearchCloud.ppt

Comparison of different server models for hosting search.

- Currently supporting 52 organisations
- 2 C/I, 5 QP s, 2 x LB (low power)
- High redundancy; High capacity for peaks

	No. Orgs	No. Servers	Tonnes CO_2
Own servers	52	104	145.6
Virtualisation	52	Say 26 (4:1)	36.4
Search Cloud	52	2 + 5 + 2 * ½ = 8	11.2

Carbon emissions comparison of three ways of supporting Funnelback service for 52 organisations: On premises solution v. hosting on VMs, four VMs per physical server v. hosting multiple clients on single physical servers – then current model. *From my presentation at the 2010 Squiz user conference in Wellington, NZ*

6.12 Life on Black Mountain

Funnelback's first accommodation was in the EpiCorp buildings off Clunies-Ross Street, part of the CSIRO Black Mountain campus. These buildings had housed CSIRO's Division of Computing Research (DCR), and then CSIRONet. The oldest part of the building originally housed CSIRO's Control Data computers: CDC3600, later CDC7600. The building was later extended to house the "Terabit FileStore" and FACOM mainframes.

Epicorp was a technology commercialisation organisation set up in 2001 by ANU, UC, CSIRO and the ACT Government. It provided space in the old CSIRONet buildings and other assistance to start-up companies in the region. Funnelback initially rented two large-ish rooms.

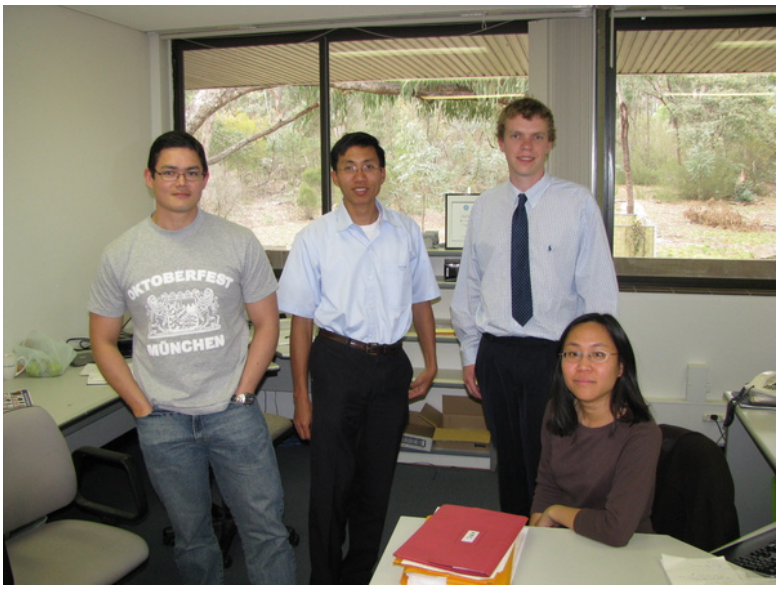
There were quite a number of start-ups accommodated in Epicorp, and a shared kitchen. The kitchen had an espresso machine but there was never any coffee for it, and its frothing wand was never cleaned. Funnelback decided to locate a rented capsule coffee machine in the shared kitchen. The coffee wasn't as nice but the machine didn't need cleaning and the capsules could be kept in our offices, rather than purloined by our neighbours. Others could use the machine, if they brought their own capsules.

After a year or two, CSIRO reclaimed the Epicorp building, using it as temporary storage for entomologists, before demolishing most of it. We were moved into a more beautiful former DCR/CSIRONet building whose seminar room had been converted in the mid-eighties to house Australia's then fastest computer – a Cyber 205.[19] We shared the building with Cambia,[20] a fascinating not-for-profit funded by royalties from patents. It had a strong interest in patents and operated the Patent Lens service.[21]

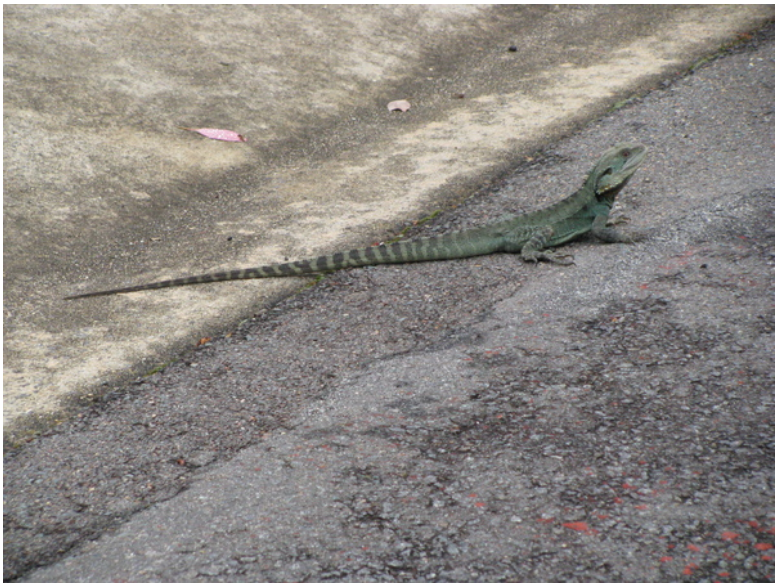
One of the enjoyable things about the Black Mountain location was the proximity to nature. We shared the bush with flocks of many different varieties of parrots and cockatoos, and eastern water dragons from the National Botanic Gardens dropped over occasionally to sun themselves in our carpark. When leaving work in the evening, we often saw kangaroos grazing between us and the CSIRO glasshouses.

[19] https://en.wikipedia.org/wiki/CDC_Cyber
[20] https://en.wikipedia.org/wiki/Cambia_(non-profit_organization)
[21] https://en.wikipedia.org/wiki/The_Lens

Cambia's founder Richard Jefferson operated a remarkable coffee machine, which I believe was a Bezzera Eagle,[22] but didn't invite us to share it. *From:* ***https://gocce-italian.com/***



2008: The Funnelback Support team in the nicer Black Mountain building. Note the lovely bush setting through the window. From left: Alwyn Davis, Puttick Hok, Jim Tink, and Kat Ng.



2008: An eastern water dragon in the Epicorp car park

[22] `http://www.bezzera.it/?p=articoli&id=6&lang=en`. Australian list prices are now around $20,000!

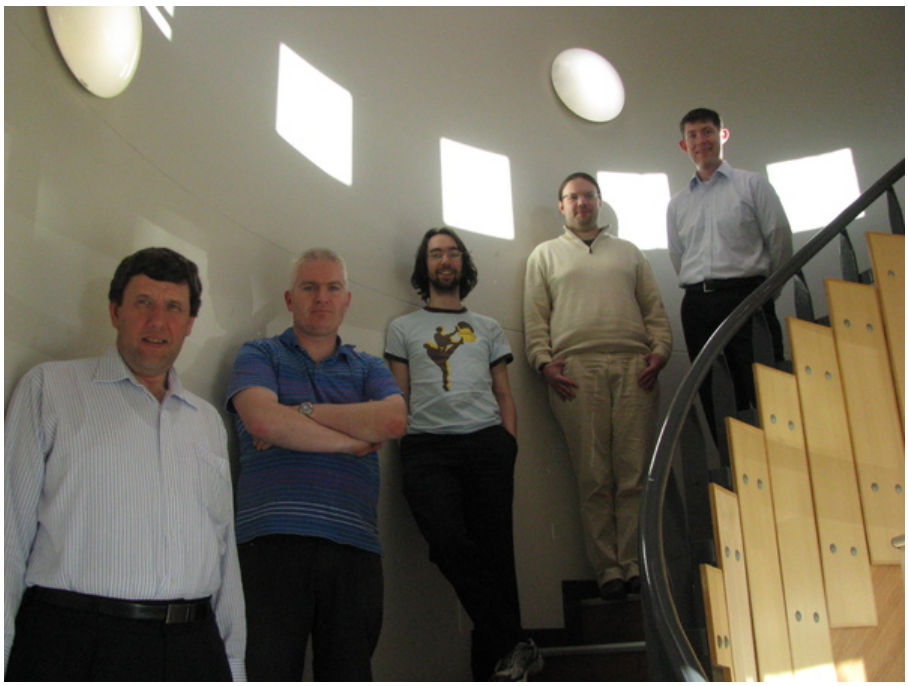
6.13 Moving to Dickson

2009: Official opening of the Funnelback office in Dickson, held in the cafe below. Jon Stanhope (ACT Chief Minister) talks to Stuart Beil. *Photo supplied by Stuart Beil.*

2009: Dan Nitsche walks through the new Dickson office at the time of the official opening,"totally oblivious" to the camera. *Photo: Stuart Beil.*

In 2009 we left our woodland premises and rented upstairs offices in Challis St, Dickson. It was above a "Cafe by Day, Restaurant by Night". The cafe was quite a good meeting spot, but unfortunately we didn't find its coffee enticing. Later when it became "Trev's" the coffee improved, and I often worked on my laptop at its outdoor tables, drinking double-shot macchiatos and sometimes eating seafood pizzas. Dickson was a major step up in eating and drinking. I recall that we sampled around 30 different Dickson establishments for lunch.

2010: How many Funnelback search experts does it take to assemble an office chair? Gordon Grace and Nalani Bakker



2011: The R&D team in the Dickson office. From left: Cliff Henderson, Francis Crimmins, Tim Jones, Nicolas Guillaumin, Matt Sheppard.

Some lunch spots were better than others. A burger and sandwiches shop around the corner was run by an incredibly hard-working woman who came in at about 5am to start making cakes and pastries. Unfortunately, due to inadequate smoke extraction, eating there entailed the inhalation of a large quantity of smoke and grease droplets. The business also practised the standard way of competing with other businesses – offering huge quantities of chips alongside an already highly calorific steak sandwich at negligible extra cost.

2009. Me, Brett Matson, Julian Cruickshank and Annie Pritchard in the Challis St Dickson office. *Photo: Funnelback*

When the original owner sold out, it was taken over by a new proprietor who took many months to convert the fit out. During the fitout customers all found new places to eat and drink. When the new business eventually opened with nice coffee and a huge range of really good cakes, no-one came. It was sad to watch over the next few weeks as cakes were thrown out, staff were let go, and the business gradually spiralled down to closure.

2010 Christmas party in New Acton. Top: Narelle Bortolin, Annie Pritchard, Nalani Bakker. *Photo from Brett Matson.*

Same Christmas party. Stuart (Santa) Beil, Francis Crimmins, Alwyn Davis, Nicolas Guillaumin. *Photo: Funnelback*

Top: March 2010 - All staff, at Challis St Dickson office. Back row: Me, Matt Sheppard, Brett Tait, Francis Crimmins, Annie Pritchard, Narelle Bortolin, Steve Barnes. Front row: Brett Matson, Ben Pottier, Cliff Henderson, Dan Nitsche, Enid Bulman, Nalani Bakker, Stuart Beil. *Photo: Maree Higgs*

We heard worrying stories of some of the eateries in Dickson: A member of staff stabbed by a co-worker in one restaurant; a cook peeling prawns while sitting on a toilet; and a kitchen hand chopping up chickens in the carpark. But there were lovely meals to be had. We often went to Cafe Praga, run by a Polish family.

6.14 CSIRO: Looking For the Exit

The board of Funnelback was very occupied with finding a lucrative but sensible exit. A very promising deal was in the offing with Brand Hoff, of Tower Software. Tower had done full due diligence and had agreed to buy the company for $5M. Three days before signing, Tower were acquired by HP and the deal fell through.

Years before, I had approached Tower to license P@NOPTIC. They embedded another Australian search engine (ISYS) in their record management system product, for an annual fee rumoured to be in the region of $0.5M. I pitched to Tower's CTO the benefits of more effective search and hinted at a cost saving. He was totally unmoved, saying that people never wanted to search the content of folders, they only wanted to locate a folder by its name. Also he said, nobody ever looks at a folder more than two weeks after it has been added to the system. I'm not sure what caused such a dramatic turnaround in their thinking.

Another possible buyer was FAST Search and Transfer, but it was acquired by Microsoft in early 2008. During discussions, FAST Australia were shocked to discover how few employees Funnelback had.

Note that the FAST technology and staff involved in supporting web search had been acquired by Overture in 2003,[23] leaving the enterprise search capabilities with FAST Search and Transfer.

According to Wikipedia[24], FAST Search and Transfer offices in Oslo were raided by Norwegian authorities in October 2008 in connection with alleged anomalies in the financial accounts.

There were many attempts to interest organisations and individuals to invest in Funnelback. Some would-be investors were quite predatory. One wanted to buy a job lot of companies from CSIRO and would have mandated a contract under which every year that each company failed to massively grow their revenue, the investor would earn a massive increase in ownership. A much more reasonable approach came from Nick McNaughton of Blue Cove Ventures, representing a Japanese investment fund and some other local investors. In preparation for substantial Blue Cove equity in Funnelback, Funnelback installed a new CEO, Jeff Pope. Late in the day, the Japanese withdrew their approval to invest and the deal fell over.

Having spent large sums on a series of CEOs (Frank Liebeskind, Jason Bresnehan, and Jeff Pope) and on external consultants, the Funnelback board decided to appoint a cheaper but inexperienced CEO in the person of Brett Matson, to be mentored by Steve Kirkby. Brett's tenure was rather longer lasting – from 2007 to 2020.

Over the years, approaches were made to major multinationals. Google had no interest in acquisition but expressed interest in the CVs of the technical employees. After Hugh Williams (whom I knew from RMIT University) started trying to recruit me to work for Microsoft, I asked him whether, instead, Microsoft would like to acquire Funnelback. After receiving my assurance that such a deal would be worth at least $10M (since Microsoft M&A would have no interest in anything less) Hugh wondered briefly whether Funnelback had enough research horsepower to serve as the nucleus of an Australian R&D office. Again, no interest in product or customers.

[23] http://newsbreaks.infotoday.com/NewsBreaks/Overture-Acquires-Two-Major-Web-Search-Engines-16748.asp

[24] https://en.wikipedia.org/wiki/Microsoft_Development_Center_Norway

My interactions with FAST

I met Knut Magne Risvik, founder of the FAST technology, at SIGIR 2000 in Athens and at the Infonortics Search Engines Meeting in the same year. He invited me to their headquarters in Trondheim to discuss web search evaluation with them and supplied me with a million web queries.

It was freezing and I had to go through Hell to get there.[a] FAST was located in a modern conversion of an ancient brewery. I had good discussions with Per Gunnar Auran.

The Norwegian airport which serves Trondheim is very close to Hell. ***This brilliant postcard – note the flames in the sky, the black beast on the tracks and the sign saying "Gods Expedition" (goods depot) – is regrettably uncredited.***

When I joined Microsoft, I found that Knut Magne was also there in Seattle, working on efficient index structures for Bing. I also visited Øystein Torbjørnsen and the former FAST technical team in Oslo as part of my campaign to improve search of email. None of the people I interacted with were implicated in alleged financial improprieties.

[a] Look up Hell, Norway on your favourite maps app.

We will see in the next chapter that Funnelback was eventually acquired by Squiz Pty Ltd. I believe that the money received by CSIRO for the sale to Squiz, plus the licence revenue received during the cottage industry phase, fully covered the cost of the research investment over the years.

In the meantime, here is some information which was relevant at the time of the exit discussions: local competition and to protection of intellectual property.

6.14.1 Other Australian Search Technology

In the early years, quite a few Australian companies were started in the general area of search. Many achieved very limited success or didn't survive for long. Among the more more successful were:

ISYS Search

ISYS[a] was a text retrieval system for Windows PCs developed in 1988 by Ian Davies. I remember receiving a CD-ROM of ISYS while working in the ANU Computer Science Department, long before becoming interested in text retrieval. (I didn't try it because we had no Windows PCs.) Over the years, ISYS developed a significant client base in the USA, mostly in local government and law enforcement.
I remember meeting Ian in Sydney with Trieu Hoang. Trieu was impressed that Ian and I executed the secret handshake known only to search engine developers. P@NOPTIC's capabilities and client base had almost no overlap with those of ISYS. It ran on Linux and provided very effective search of web sites and its client base was mostly universities and government agencies. ISYS had a much larger library of filters for different document formats. In the end, nothing came of the meeting.

[a] `https://en.wikipedia.org/wiki/ISYS_Search_Software`

WebWombat – Searching the Australian Web

WebWombat was founded by Rod Ashcroft and Phil Bertolus in 1995. It aimed to provide search of web sites in the `.au` domain for Australian searchers. Phil Bertolus was the CTO and he gave a keynote talk at an Australasian Document Computing Symposium. He told how the company started with its servers in a house in a Melbourne before moving to an office in St Kilda Rd, and growing to a maximum index size of around 30 million pages.
WebWombat was funded by advertising and site promotion. According to Phil they hadn't invested much in the science of result ranking. Indeed he said that they sometimes deliberately inserted irrelevant but interesting results – they found that people searching for lawn mowers or tax agents were quite likely to click on results for holidays in Tahiti. ☺

LookSmart

LookSmart was founded in Melbourne by Evan Thornley and Tracy Ellery. It was initially called HomeBase and was a companion portal to the Reader's Digest. In 1998, it launched the LookSmart search engine. It operated rather more like the Yahoo! directory than a standard web search engine. It was very heavily visited and for a time very financially successful. It was contracted by Microsoft to provide directory and listing services for five years from 1999.

6.14.2 Protection of Intellectual Property

In CSIRO, I adopted the policy of protecting IP by publishing it. That way it couldn't be patented against us. The value to CSIRO of patenting our discoveries didn't seem to justify the very high cost of gaining, maintaining, and particularly defending patents.

The Funnelback board thought that the value of the company would be increased if our IP was protected. I was willing to oblige, but had misgivings. It would be very difficult to detect whether competitors were violating a patent, and Funnelback did not have the money to defend a patent against a large multinational competitor.

To patent something we needed ideas that hadn't been previously commercialised or published and would stand up to *prior art*[25] attack. That ruled out most of our main product, but the ideas behind ANNIE seemed to satisfy the criteria and we filed a patent application.[26]

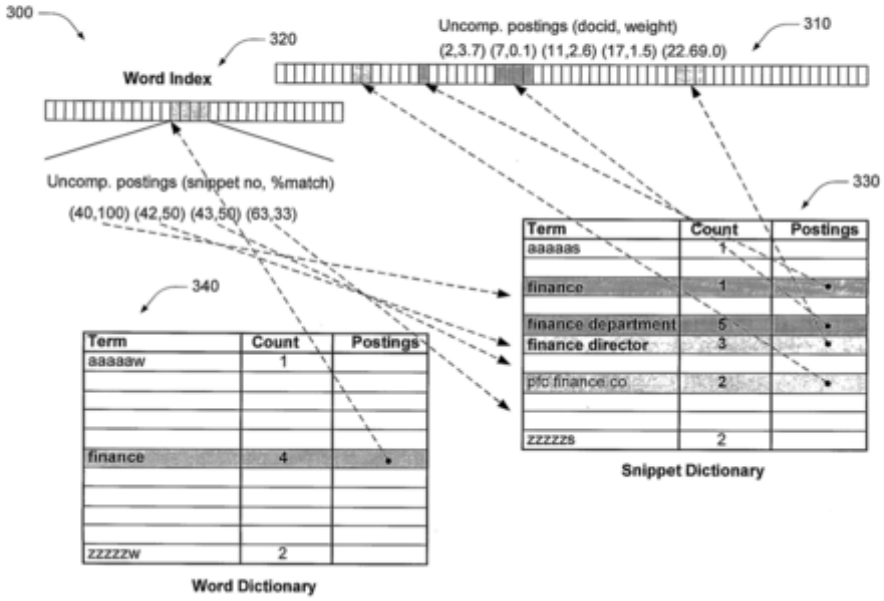
[25] In patent law, *prior art* means evidence from commercial use, publications, or patents that your invention is already known.

[26] `https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2008061290`

As outlined in Section 4.13.1 on Page 78, Tom Rowlands and I had been studying the ability to retrieve documents based on textual annotations external to the documents. In 2007 I travelled from Glasgow to Aberdeen for discussions with a research colleague Ayse Goker about the company Ambiesense which she formed with her Norwegian husband. They had developed a localised information dissemination system which used hardware gadgets they designed and built. We walked around the streets of Aberdeen and as we passed historic buildings fitted with Ambiesense gadgets, Ayse's phone would receive a text providing information about the building and its history. Ayse was trying to build up the company, and wanted to hear about the Funnelback experience.

Patent: Annotation Index System and Method

[EN] ANNOTATION INDEX SYSTEM AND METHOD
[FR] SYSTÈME ET PROCÉDÉ D'INDEX D'ANNOTATIONS



Abstract

[EN]
A method of encoding an inverted list structure of annotation material is taught. This method includes the steps of: [a] collecting a group of documents; [b] determining a group of annotations referencing the group of documents; [c] forming a snippet index by grouping the group of annotations by unique annotation identifier; and [d] forming a snippet dictionary which, for each unique annotation identifier, indexes a corresponding position in the snippet index for the group of annotations having that unique annotation identifier.

On the train to and from Aberdeen I worked on the design of an efficient indexing and querying system based entirely on annotations. It involved a two level index structure. A word index referenced annotation snippets rather than documents and a snippet index referenced documents. The motivation was to reduce search latency while hopefully generating more precise answers. I soon built an indexing and retrieval system based on this idea and called it ANNIE – ANNotation Indexing Environment or some such.[27]

[27] When Annie Pritchard joined Funnelback, I wondered if she would be annoyed by the name, but she said she wasn't.

ANNIE did indeed deliver very low latency search results but in the environments we tried it, there were serious gaps in coverage. We could only achieve the needed result quality by combining ANNIE with PADRE. Running both systems for every query naturally increased latency. Using a decision mechanism to run ANNIE and then decide whether to run PADRE reduced the harm to throughput but increased the variability of latency. That's a negative for user experience.

Patent: Search Result Sub-Topic Identification System and Method

(EN) SEARCH RESULT SUB-TOPIC IDENTIFICATION SYSTEM AND METHOD
(FR) SYSTÈME ET PROCÉDÉ D'IDENTIFICATION DES SUJETS SECONDAIRES DANS UN RÉSULTAT DE RECHERCHE

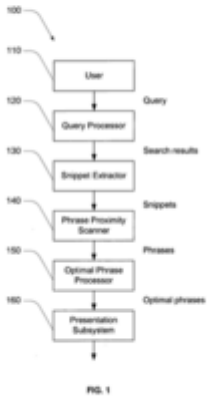


FIG. 1

Abstract

(EN)

A method and apparatus for sub-topic identification from a search result that matches a query, said method including the steps of receiving a search result, extracting snippets from said search result that contain said query, truncating snippets on an instance of a boundary token, identifying phrases within said snippets that include the query, comparing all said phrases to determine optimal phrases, and presenting said optimal phrases. The apparatus for sub-topic identification from a search result that matches a query may include a dedicated server or a proxy for processing the search and sub-topic query.

Brett's Fluster Patent

Soon after the ANNIE patent, and for the same reasons, another patent application[28] was lodged for Brett Matson's Fluster query refinement system. (See Page 162.)

It cost hundreds of thousands of dollars to file and maintain these two patents, and in the end they played no role in persuading Squiz to purchase Funnelback. They turned out to be an investment of rather dubious value.

Later, while working for Microsoft, I found a very different patent situation. An in-house legal section, highly experienced in IP matters, made the process of drafting the patent much less onerous for the inventors. Once a patent was filed, the inventors received a commemorative granite cube and a small financial reward. When the patent was granted, a wooden plaque was given, though I left Bing before I could confirm that. The office of Eric Horvitz, Head of Microsoft Research, contained almost enough granite blocks and wooden plaques to build a house. ☺

[28] https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2008098282&_fid=US43513252

Chapter 7

The Early Squiz Years

Squiz Pty Ltd was a company formed in 1998 by John-Paul Syriatowicz (JP) and Steve Barker. JP had developed a content management system (CMS) called Matrix which he open-sourced. Steve had a background in sales and publishing. Squiz's early marketing tag line was "supported open source". Unlike many CMS companies, Squiz managed to survive the dot com crash around 2000, and successfully took on many former customers of their failed competitors.

Our first contact with Squiz was a lunch meeting with JP at Vivaldi's on the ANU Campus, organised by Stuart Beil, quite a while before the Funnelback spin-off. Nothing came of the meeting, but perhaps a seed was sown.

Brett Matson remembers discussions with JP:

> ... the discussion being had with JP was that he'd offered Funnelback a flat fee for the ability to give Funnelback to all Squiz customers. Cannibalising an unlimited amount of our core market for a fixed amount sounded like a terrible deal to me and JP wasn't willing to increase his offer, so the conversation went nowhere. I guess a flat rate was the only thing that made sense to Squiz given its open source business model and therefore its reluctance to charge a licence fee.

7.1 How the Squiz Buy-Out Came About

Skipping forward several years to 2009, Funnelback was contacted by Squiz UK, a Squiz Group subsidiary managed and partly owned by Steve Morgan, who had been Steve Barker's brother-in-law. Squiz UK had a number of prospective deals which they looked like losing through lack of an effective search capability. Was there a way that Squiz and Funnelback could work together? At the time there seemed to be no keenness on the part of either JP or the Funnelback board to talk about a merger or acquisition.

I was scheduled to travel to the UK for another reason, and gained approval from Stuart and Brett to meet Steve Morgan and discuss the possibility of our providing Funnelback as the search solution on an opportunity-by-opportunity basis.

I arranged to meet Steve Morgan in what turned out to be the ramshackle Squiz UK offices on the top floor of an old building in Clifton St, Shoreditch in London's East End. Shoreditch had aspirations as a high-tech hub and started branding itself as *the Silicon Roundabout*. Steel and glass buildings full of tech and digital design companies were gradually spreading out from Liverpool St Station. Steve had negotiated a long-term lease on Clifton St at what became a very good price.

Lo and behold, there was another person already in the meeting room: JP! Steve Morgan thought that there would be merit in Squiz acquiring Funnelback and arranged for JP to be there. From memory, I pitched the merits of Funnelback search, and showed demos we had (of course) set up for the clients we knew Squiz UK were hoping to win. Then a discussion ensued about whether CSIRO would be willing to sell Funnelback – yes. JP then asked me what would be the likely asking price. I (honestly) said that I didn't know but that I thought that CSIRO would prefer a buyer with an interest in using and developing Funnelback technologies.

Discussions proceeded rapidly between CSIRO and Squiz. Squiz agreed to buy Funnelback with

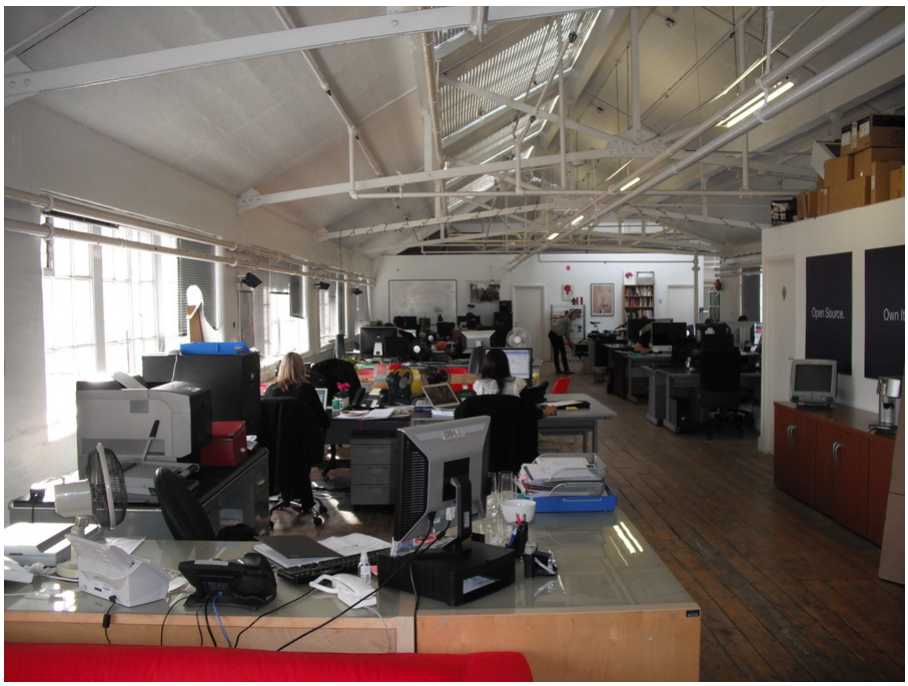
vendor-finance over three years.[1] The former Funnelback board was dissolved and a new board comprising JP, Steve Barker and Stuart Beil was consitituted. Options held by the founders were cancelled and arrangements were made to pay the holders the appropriate percentage of each payment made by Squiz to CSIRO. Since the strike price of the options granted to me when I became a Funnelback employee was much more than the value of the shares implied by the purchase price,[2] I was sent a cheque for one cent, in compensation for the cancelling of my options. I complained to Karl Rodrigues that they should have given me the chance to exercise my options. He said, "Send the cheque back if you want and I'll send you an invoice for the tens of thousands of dollars!" ☺

From the point of view of the Funnelback team, the initial period of Squiz ownership was a halcyon time. For a start, it ended the considerable distraction and time spent in pursuing possible acquisition or investment deals. Furthermore Squiz encouraged us to do what we did best and generate as much profit as possible. They said, "if you have to partner with Squiz competitors to make money, by all means do so." During this era, the board and the team were very well aligned.

2009: Zetland House in Clifton St, Shoreditch EC2A 4HJ. The building orginally housed a note printing works for the Bank of England. The original Squiz UK office was on the very top floor. Steve Morgan eventually leased one of the parking spots seen here in order to park his BMW Z4 sports car.

[1] In the event, Squiz extended the period to five years. I believe that the debt to CSIRO was paid off entirely using dividends from Funnelback.

[2] Under Australian law, if I had been given a strike price lower than the value of shares, I would have been immediately liable to pay tax on the difference, even though I had no way of selling the shares until CSIRO exited.



Inside the original Squiz UK office at a remarkably quiet time. ***Photo: Stuart Beil***

2013: Lisa Caswell at work in the old Squiz UK office. She later married Steve Morgan.

7.2 Early UK Projects

In mid 2009, the acquisition had happened and I was back in London again, hanging out in the Squiz UK office and presenting at a Squiz industry summit held in the magnificent surroundings of Australia House on the Strand. The event was well attended and seemed very successful. It was followed by drinks for staff and customers on a boat moored on the Thames. Some people drank rather heavily and one dark-skinned, dark-haired woman, dressed in black, insisted on walking in the dark along the embankment road, with cars whizzing past.

2009: Setting up for the Squiz event in Australia House.

I found a big difference in attitudes to after-work drinks in Canberra and London. In the Squiz Office, staff frequently repaired after work to The Fox, or to the Old Kings Head. That almost never happened in Canberra with occasional exceptions on Friday evenings. When I thought about it, the explanation seemed clear. In Canberra, nearly everyone drove a car and was keen to stay below the blood-alcohol limit and to get home as soon as possible. In Shoreditch, no-one drove, and spending an hour in the pub meant that public transport would be a little less crowded. Squiz UK staff spent ages (and lots of money) commuting. Two lived on the south coast (at Brighton and Worthing), one came in from Reading, one came from Horsham, and another from Caterham.

Commuters from Sussex loudly complained that they paid thousands of pounds a year for their rail passes, but when they boarded the trains they found every seat taken. The CEO of the railway company told them that they should move to the coast and get on at the first station!

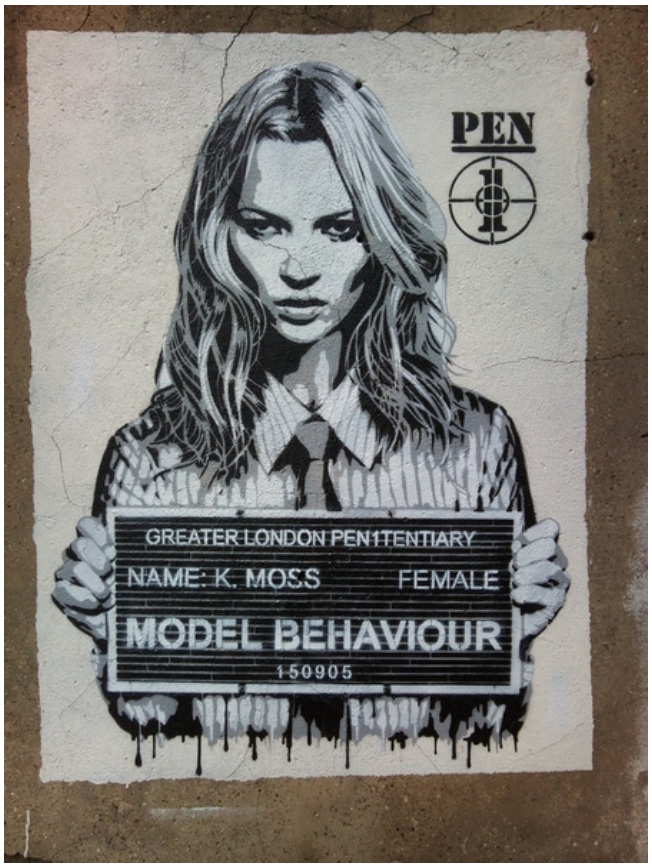
To constrain costs I stayed at Marlin Apartments in a yet-to-be-developed area of Canary Wharf. I was able to use a very relaxing method of travelling to Australia House, cruising up the river on the Thames Clipper service – almost door-to-door. In many later visits I stayed at the even cheaper Marlin Apartments in Stratford, watching slums and ruined old factories gradually making way for an enormous Westfield shopping centre and facilities for the 2012 London Olympics. I got a taste for what it would be like being a commuter in London. I didn't relish it, but felt pleasure in mastering the London transport system, and knowing both the best ways to get places and how to cope when stations closed or trains broke down.

Stratford has an enormous station accommodating mainline trains from Essex, the high speed Javelin trains from Kent to St Pancras, Docklands Light Rail (DLR), the Jubilee and Central underground lines, and the London Overground. To get to Shoreditch I had the choice of traveling to Liverpool St station either by the Central line or by the main line. I pretty soon learned that the main line had no intermediate stops, got there in half the time, and the carriages were half empty – because commuters changed at Stratford for other services into the city. An important learning experience on my way to transport enlightenment came when a local advised me to go to Marymount main line station because it was much quieter than Stratford. That was very bad advice.

First a Marymount station anouncement told me, "National Rail would like to apologize that the 08.24 for London Liverpool has been delayed by 12 minutes." Since trains ran every 12 minutes I would have had no way of knowing of the delay if they hadn't told me! When the train pulled in at

08.36 the doors opened and clouds of heat and steamy fug billowed out. Commuters were crammed in and no-one could move. A couple of would-be passengers unsuccessfully tried running at the wall of people before bouncing off and giving up. "National Rail would like to apologize that the 08.36 for London Liverpool has been delayed by 12 minutes." When I eventually managed to squeeze onto the 09.00 at 09.12, I found that two-thirds of the passengers alighted at Stratford and after that I had a group of seats to myself. After that, I caught the train at Stratford.

Over the years, I spent quite a lot of time in the Squiz UK office, one week at a time. Some of the trips were paid for by Squiz UK, some by Funnelback, and some by academic and government institutions who funded me to participate in conferences or meetings, or to participate in PhD examinations. In addition to interacting with FBUK technical staff and trying to help them feel connected to Funnelback HQ back in Canberra, I went with sales and project staff to meetings all over the country, and presented at conferences.



On one of my early visits to Squiz UK, this Banksy artwork appeared on a wall somewhere near the office.

One memorable early project visit was to Dyson HQ in Malmesbury, Wiltshire. I travelled with Lisa Caswell and a Squiz implementor. It brought home to me how different the essential business models were between Squiz and Funnelback. Rather than the lightweight projects favoured by Funnelback, the Squiz Dyson project took the full project management route, with meetings, agendas, specifications and reviews. It was quite cool to see the display of innovative Dyson products: stylish vacuum cleaners, heaters, fans and hand dryers.

On one occasion, there was an opportunity for Funnelback to provide the package holiday search for the UK's biggest leisure travel company TUI Travel, which operated in 180 countries. The particular focus of the project was on Russia. There were significant challenges for us in meeting the requirements of how search should work. TUI demanded that the search should never produce an empty results set. That wasn't a problem, but the issue was how to rank the approximate matches. If you asked for a mid-price holiday 11–18 June to Sharm-el-Sheikh, departing Moscow, and no such holiday were available, would you be more likely to take up a holiday at exactly the same time, to another Egyptian destination like Luxor, or to a swimming holiday destination elsewhere in the Middle East? Or would you prefer a Sharm-el-Sheikh holiday for a shorter duration, or a longer one, or

the same length at a different date. Or a cheaper one, or a more expensive one? User interaction data might allow the best ranking to be determined, but we had none.

The TUI search also had to respond "instantly" to changes in inventory. Package holidays had to disappear from the search as soon as all seats were sold.

I flew over to the UK with Stuart Beil and spent time on the flight learning the Russian alphabet, pre-processing and indexing the test data set we'd been given, working out how to best handle fast updates and instant deletions, and coding needed extra capabilities, like Russian spelling correction, in PADRE. The day before meeting TUI, we had other sales meetings in towns outside London. I remember continuing the TUI coding on trains and on station platforms.

Finally, we took the train to Coventry and presented to TUI. We made a credible presentation and I think convinced them that we could do the job. However, we didn't get the gig. ☹

Another exciting e-commerce opportunity was with the UK's largest supermarket chain, Tesco. Again there were significant technical challenges. First, search results had to be restricted to stock items available in the searcher's local area. Second, the search interface had to be able to work in the Czech Republic. One of the quirks of the Czech language is that, when sorting, the two-character combination `ch` is treated as a single letter, and it sorts after `d`. After a fair bit of coding, and some work on a Tesco demo, we persuaded the Tesco person we were dealing with that Funnelback was the tool for the job. Unfortunately, the board of Tesco decided to include the search capability in a very large higher-level contract with a multinational, perhaps Oracle. This was very disappointing but Stuart Beil wasn't surprised. He said that we had failed to identify the key decision makers and to persuade them.

Looking back I realise that I was involved in many unsuccessful pitches. We failed to persuade Lloyds Bank, the Royal Courts of Justice, Goldmark Jewellers, Skyscanner in Edinburgh, a Scottish government agency, and a defence contractor in Basildon.

We went to Basildon with a representative of a multi-national IT integrator. The client had tens of terabytes of documents on a fileshare and a much larger quantity archived in a data vault. The task was to provide effective search over both the fileshare and the vault. The integrator told us that if we had a solution for this, they would immediately take us around dozens of other clients with similar problems. Unfortunately we discovered that the vault vendor placed insuperable obstacles, both technical and contractual, to indexing the content of the vault. Much later I was told by David Sitsky, Chief Scientist at Nuix, that Nuix solved this problem by ignoring the vault product and accessing the underlying database directly. Because their remit was legal discovery, unlike us, they didn't need to enforce document level security.

On a more positive note, I was involved in more successful visits to the Victoria and Albert Museum, to the Royal College of Nursing, and to the British Medical Association. We were also particularly successful in the university sector. I've previously mentioned that P@NOPTIC acquired the University of Staffordshire as a customer quite early on. Also before the Squiz acquisition, Stuart Beil managed to sign up the London School of Economics and Political Science (LSE) as a client.

LSE had deployed a Google search appliance, but didn't brand it as such. To my delight, the LSE Head of Web Services, Stephen Emmott, told us that they received complaints from visitors to their site, along the lines of, "Your search is hopeless. Why don't you just get Google?" Stephen told me that LSE maintained a list of more than 600 "key matches". A key match is a short cut of the form: *query* $\rightarrow$ *best_answer* which causes the appliance to present *best_answer* at the top of the results ranking when *query* is received. It's an admission of failure of course but all search engines, including Funnelback, include similar mechanisms.

LSE's concern was with the significant staff cost of maintaining the key matches file. They were also interested in using Funnelback to support expertise finding and in the `Fluster` technology [3] developed by Brett Matson. Fluster was a tool for suggesting query refinements to searchers, based on linguistic contexts. For example the query `policy` might suggest refinements such as `foreign policy`, `work from home policy`, or `policy on stress leave`.

[3] See Page 156 for information about the Fluster patent.

2009: LSE Head of Web Services, Stephen Emmott (left) and me. We had just been to lunch at Britain's deepest pub. *Photo: Stuart Beil*

In March 2009, I met Stephen Emmott and his colleague at LSE in Houghton St, very close to Australia House and Temple Bar. Stuart and I met him again in November that year. Stephen owned a fold-up Brompton bicycle. Each day he rode his Brompton to the train, folded it up on the ride to Marylebone station, then unfolded and rode to LSE.

Stephen and LSE were exacting clients – holding us to account on response latency, update speed and result quality. They also got Funnelback for a good price, which took into account their value as a reference site. Stephen was fully prepared to present LSE's experiences at Squiz/Funnelback user conferences, and one year we paid his fare to present at a user conference in Australia.

7.3 ∗The Emmottiser

Two important Funnelback features owe their origin to Stephen. At the Australian conference he reported his troubles at LSE when an academic would come complaining that his web page W should rank highest for queries relating to his field of expertise (say `omphaloskepsis`), but instead it didn't appear, or ranked below one of his rivals! What Stephen wanted was a automatic tool which would say, Your page doesn't rank top for `omphaloskepsis` because:

- It's not in the crawl because it isn't linked from anywhere, or;
- It doesn't contain as many occurrences of `omphaloskepsis` as your rival's, or;
- `omphaloskepsis` is not in the title, or;
- People searching for `omphaloskepsis` click on your rival's page rather than yours, etc.

With encouragement from me, Tim Jones from our R&D group set about building what I called *the Emmottiser* but which was eventually given a more marketable name by the company. It wasn't a tool you could give to academics because of the danger of triggering a harmful internal SEO war, but something which could be run by the search master to give advice to academics.

7.4 *FineTune

The second LSE-requested feature was *FineTune*, a tool for tuning search rankings. Out of the box, the Funnelback ranking at LSE left something to be desired. The PADRE query processor had many knobs and buttons which could be adjusted to improve things (or make them worse ☺). Humans tend to jump to a conclusion based on one or two examples and in response fully rotate one of the knobs. For example, "Upweight phrase matches and matches in titles", "Downweight matches in URLs", "Stop using anchor text", "Rely on click data." Experience showed that dramatic changes usually fix one or two problem examples but make things worse overall.

In contrast, an algorithmic tuner can optimise performance across a large set of test queries. If the test queries are representative it is reasonable to expect that results for queries not in the test set will also improve. Importantly, if the test queries are representative of the total workload then improvements on the test will manifest as improvements in overall user experience. This is the reason to use the set of queries and right answers for tuning, rather than using it as a key matches file.

I built a tuner called *FineTune* which tuned one knob at a time. It found the optimum setting for that knob, by moving it in small increments and re-running the test queries for each setting. Because the knobs interact with each other, multiple passes are needed. The process continues until no further improvement is achieved.

FineTune made use of the formats defined in *C-TEST*, the CSIRO Toolkit for the Evaluation of Search Technology. A C-TEST file is in XML format and contains a `query` element for each test query. Query elements then contain `interpretation` sub-elements. Each interpretation then contains `answer` elements, each including the URL of a useful answer and a utility score for that answer. (C-TEST is described further on Page 65.)

For example the query `jaguar` might have two interpretations: `car` and `animal`. For the car interpretation `jaguar.co.uk` would be an answer with high weight, and for the animal interpretation, `en.wikipedia.org/wiki/Jaguar` would be an answer with high weight.

At LSE, Stephen Emmott and his staff manually created a C-TEST file, using more than 100 "business critical" queries such as `enrolment`, `scholarships`, and `environmental studies`, and answers they found by searching, browsing, and from their own knowledge. FineTune worked on this file and made a dramatic reduction in the number of failed queries – queries for which no useful answer appeared in the top ten results. The automatic tuner achieved far more improvement than humans were able to achieve.

FineTune output included a large table in which rows (corresponding to queries) were sorted by the utility score achieved. By looking at the bottom rows of the table you could see the failures. Sometimes these failures would point to a problem in the creation of the C-TEST file. In others, the failure would indicate a lack of a suitable answer page. At LSE, gaps in information published for students were identified and new content created. In some cases, there would be a case for adding a key match or a query rewriting rule. For example, on a government site, the query `car rego` might need to be mapped to the page on `motor vehicle registration`.

Cells in one column of the FineTune table would list the URLs identified in the test file as useful for the relevant query. If a URL was greyed out, that meant that that URLs was not in the index. Perhaps the inclusion and exclusion rules for the crawler needed adjusting. Perhaps there was a redirection which needed to be taken into account.

In another column was a link to rerun the test query, allowing the search master to see a side-by-side comparison between the tuned and untuned versions of the results. Sometimes this revealed useful answers deserving of being added to the test file.

Matt Sheppard turned FineTune into a very nice feature accessible from the Funnelback Admin-

istrator page. Its interface allowed you to type in a query, or randomly sample one from the accumulated query log (the scientifically preferred approach). You could then find and add useful answers by internal or external search, or by browsing. Once a query was entered the system would run it and evaluate the results, allowing you to correct any errors.

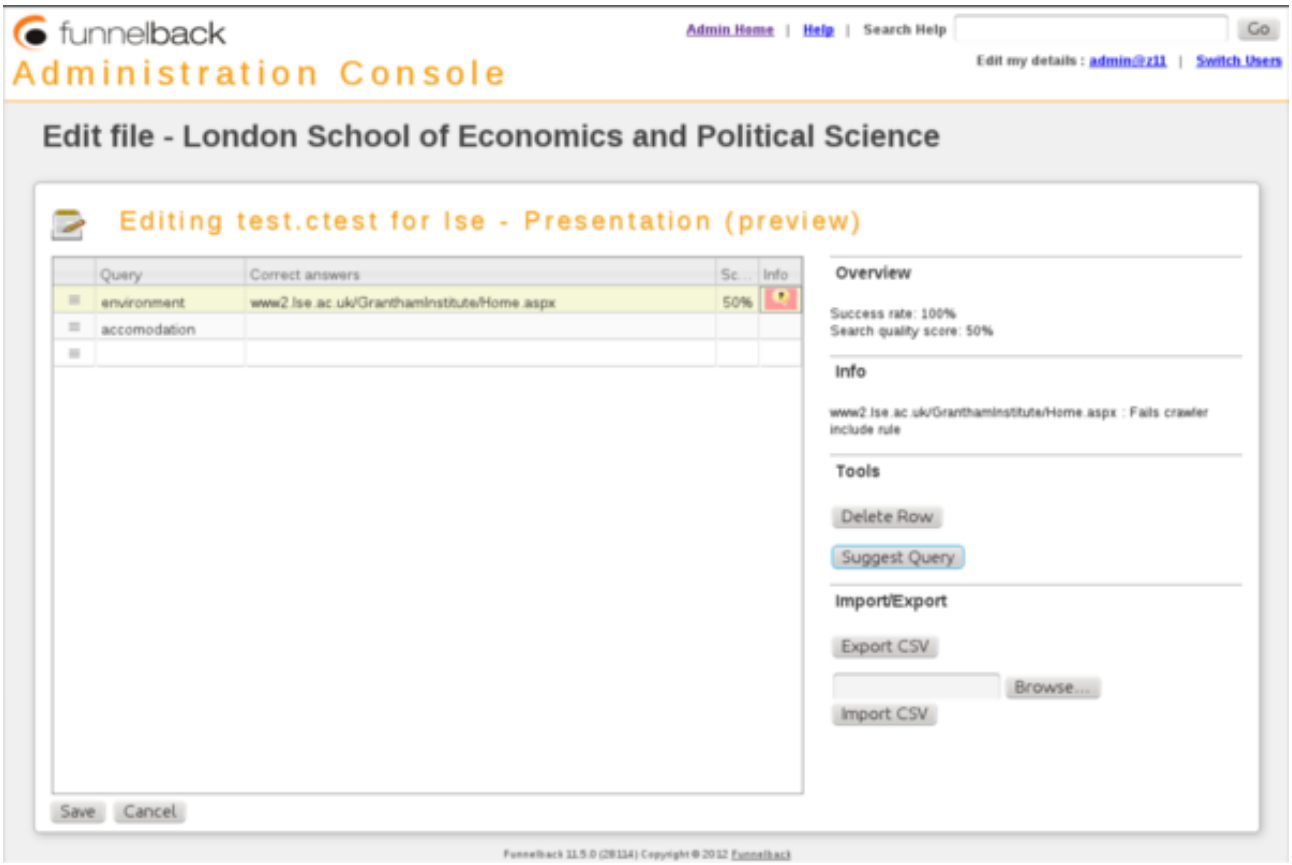
After setting up a testfile in this way, you could press a button to initiate a FineTune run. Once tuning was complete, the search master would be alerted and could check the relative performance of the tuned settings before clicking a button to push them into production.

Out of scientific interest, I converted the Google key matches file from the old LSE installation into a C-TEST file. After tuning with the key matches test file, error rates were higher than for the manually constructed file. This was not surprising since these were difficult queries, ones which the Google appliance had done badly on.

I also created a C-TEST file from the clicked queries log file – in which the best answers for a query were the URLs which most people clicked on. Utility scores were calculated from the proportion of clicks each answer achieved. For this test, failure rates dropped to zero and results neared perfection. FineTune learned that the weighting of click data in the ranking should be as high as possible. This is not good since the test set is highly unrepresentative of the overall workload and gives no guidance on how to best process the many queries for which there is no historical click data.

Building the testfile from randomly sampled queries should do the best job at tuning for overall user experience. However, it is arguable that tuning on business critical queries will best achieve the aims of the organisation.

Developing a method and supporting tools through which enterprise search customers could significantly improve the quality of their search with a minimum of time and effort was an achievement of which I am quite proud. Matt Sheppard's productisation of FineTune within the administration interface made it even better and much more useable.



2012 version of the tuning section of the Funnelback admin interface.

7.4.1 *The Wrong Way to Tune Search Results

Later, at another major university, a consultant was managing the new Funnelback installation project. The consultant called frequent project meetings and each time they raised queries which were not producing the desired results, maintaining a list of problem query Action Items. "Well you guys have fixed the undergraduate scholarhips problem but now there's a problem with environmental mechanics. As each query was fixed, the query was dropped from the list of action items. This process was potentially never ending, since manually tuning for one query was likely to break others. It meant that every meeting was focused on Funnelback failures, and to me seemed like a make-work exercise.

I attended one of the meetings and argued, I think successfully, that while individual problem queries were important, it was the overall performance that mattered. We didn't want to shy away from problem queries – each of them (even the ones previously "solved") should be added to a testfile like the one developed at LSE. Problems remaining after tuning should be analysed and dealt with, as was done at LSE.

7.4.2 *Generalised University Testfiles

There is a huge amount of overlap across universities on the query sets which are of greatest importance to the university and to visitors to its web sites. jobs, enrolment, scholarships, research, accommodation, examinations, graduation, timetable, all feature prominently in the query load of all the English speaking universities I've looked at. Then there are subject queries such as physics, environmental mechanics, British history, and women's studies.

Based on these patterns, it was easy to build C-TEST files for universities, even without access to their query logs, and to build tuned demos when pitching to potential university customers.

7.4.3 University Course-Finders

The financial health of many universities depends upon attracting student enrolments. Prospective students visiting a university web site need to be able to quickly find out whether that university offers a course they are interested in and to be able to see outcomes from it, such as average salary on graduation, student ratings, and average months from graduation to employment.

Squiz UK had built course finders for several universities. I think that initially they were built without Funnelback, but later Funnelback was used. The Squiz business model was to custom build each coursefinder solution and charge on the basis of billable hours.

7.5 Other UK Clients

In this era the UK Government operated three major portals to its online content and services: gov.uk, business.gov.uk, and NHS Choices. Each of them provided what seemed to be rather ineffective search. Given the success we'd had with whole-of-government search in Australia, I thought that we had a good chance to improve things in the UK. I eventually managed to line up a meeting with Peter Jordan who "ran" the gov.uk portal. He was aware of our AGOSP projects, had contacts in AGIMO, and seemed interested in what I had to say. However, he told me that the three portals were contracted out to major service companies, and that it would be some time before they were re-tendered. He also pointed out that the UK government had recently legislated a preference for open source software.

2009: The highly salubrious (slightly ironically) headquarters of the UK Ministry of Work and Pensions. I went there to meet Peter Jordan who was in charge of search on the main UK Government portal.

The British Medical Journal was a Funnelback client. We went to project meetings in this impressive building. There is another plaque further along the wall, in memory of the 13 people who died when a bus was blown up outside this building in 2005. Medical staff in the BMA building rushed out to render assistance.

I also met with a senior person at Serco who had responsibility for the government business portal. His office was in a glass building on the bank of the Thames near London Bridge. We went to lunch at the Butlers Wharf Chop House, where he explained to me how the system worked, pointing out that Serco were also responsible for maintaining and operating HMS Belfast anchored in the Thames just over from us.[4]

I wasn't able to get anywhere but you will see in Sonia Piton's panel on Page 178 that Funnelback did eventually gain NHS Choices as a client. I was delighted to hear that!

[4] And also for operating some of Australia's migration detention centres!

Since I only visited the UK occasionally, I'm not aware of all the non-university projects that were undertaken. I have mentioned Dyson, and the Scottish Commission for the Regulation of Care elsewhere.

7.5.1 IWMW: The Institutional Web Management Workshop

Since 1997, well-attended IWMW workshops[5] have been held at university campuses around the UK. The first was held at Kings College in London.

Web masters at UK universities have a history of cooperating with each other, sharing knowledge, and joining forces to negotiate consortium deals with vendors. This occurred even when universities were competing fiercely with each other – If I remember correctly, billboards appeared just off the campus of the University of Bedfordshire encouraging students to enrol at a competitor. A large majority of UK universities had standardised on a content management system, Terminalfour I believe. The vendor had lowered its prices to enter the university market and in return had gained a large number of customers, and a thriving pool of customer expertise.

It seemed that this would be a great model for Funnelback UK to pursue. In 2011, I successfully proposed a talk to the two-day IWMW at the University of Reading. Matt Taylor, Phil Widdop and I made the trip to Reading. Matt and Phil staffed a Funnelback stall (near the Google one) and I presented. Unfortunately, the program for IWMW noted that financial pressures on universities had forced a loss of jobs, reduction in budgets, the shortening of the workshop program to two days, and a substantial reduction in the attendance fee. However, at the workshop dinner, I had encouraging conversations with representatives from a number of universities, including Helen Varley Sargan (Cambridge) and Dawn Ellis (University of Edinburgh).

Squiz/Funnelback UK subsequently exhibited at other IWMWs.

7.5.2 UK University Clients

Matt Taylor and I presented to both Birmingham and Manchester universities as part of FBUK's response to large search tender opportunities. In both cases we were unsuccessful. Autonomy and Verity were very strong competitors.

A significant early success in the university sector was when we signed up Cambridge University in late 2012 after presentations to Helen Varley Sargan and Jon Warbrick. First we built a proof of concept and then it went into production. I remember being taken for a drink at the Eagle, the pub famous for being frequented by James Watson and Francis Crick, the discoverers (along with Maurice Wilkins and Rosalind Franklin) of the DNA double helix.

As noted in Gordon Grace's recollections on Page 180, Phil Widdop made endless calls to university searchmasters across the UK. When they told him that they operated a Google Search Appliance (GSA), he asked them when the support period for it expired. GSA's were notorious for having only a two-year support period. If your GSA failed after that, you needed to buy a new one. On the date of a GSA support expiry Phil would phone up the owner and make the Funnelback pitch.

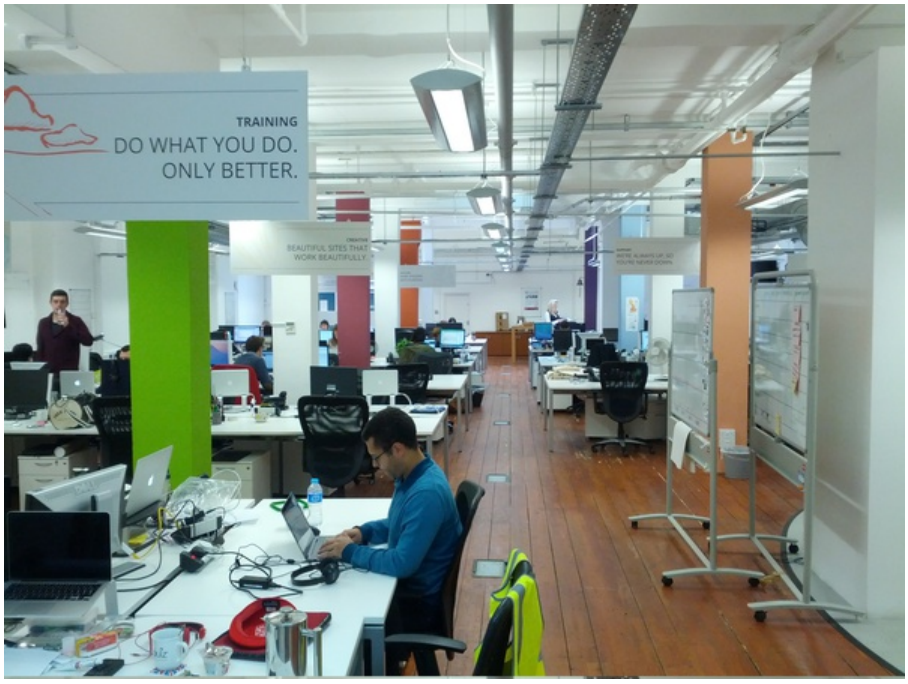
The task of recruiting GSA customers became even easier when the Google Mini was discontinued in 2012 and the GSA in 2016.[6]

I believe that Funnelback eventually acquired more than 50 UK universities as clients, including top-ranked Oxford, Cambridge, University College London, and Edinburgh.

[5] https://iwmw.org/

[6] https://en.wikipedia.org/wiki/Google_Search_Appliance

2010: Matt Taylor after our (unsuccessful) attempt to win a big contract with the University of Birmingham.



2014: The second (underground) office of Squiz/Funnelback UK, still in Clifton St, Shoreditch. *Photo: Nicolas Guillaumin*

7.6 Squiz/Funnelback Events in London

Squiz organised quite a number of events in the UK. Hannah Cooper was in charge of marketing and produced some excellent Funnelback marketing materials: pens, brochures, banner signs, and conference bags. At the event held in the conference centre at Westminster Abbey, Stephen Emmott arrived on his Brompton folding bicycle and gave a presentation using an amazing tool in which the presentation seemed to be a giant two-dimensional map on which you could pan and zoom to show what you needed to – no linear progression through a set of slides! On leaving the conference centre, we were greeted by a man dressed in colourful robes, who asked us how our meeting went. I was

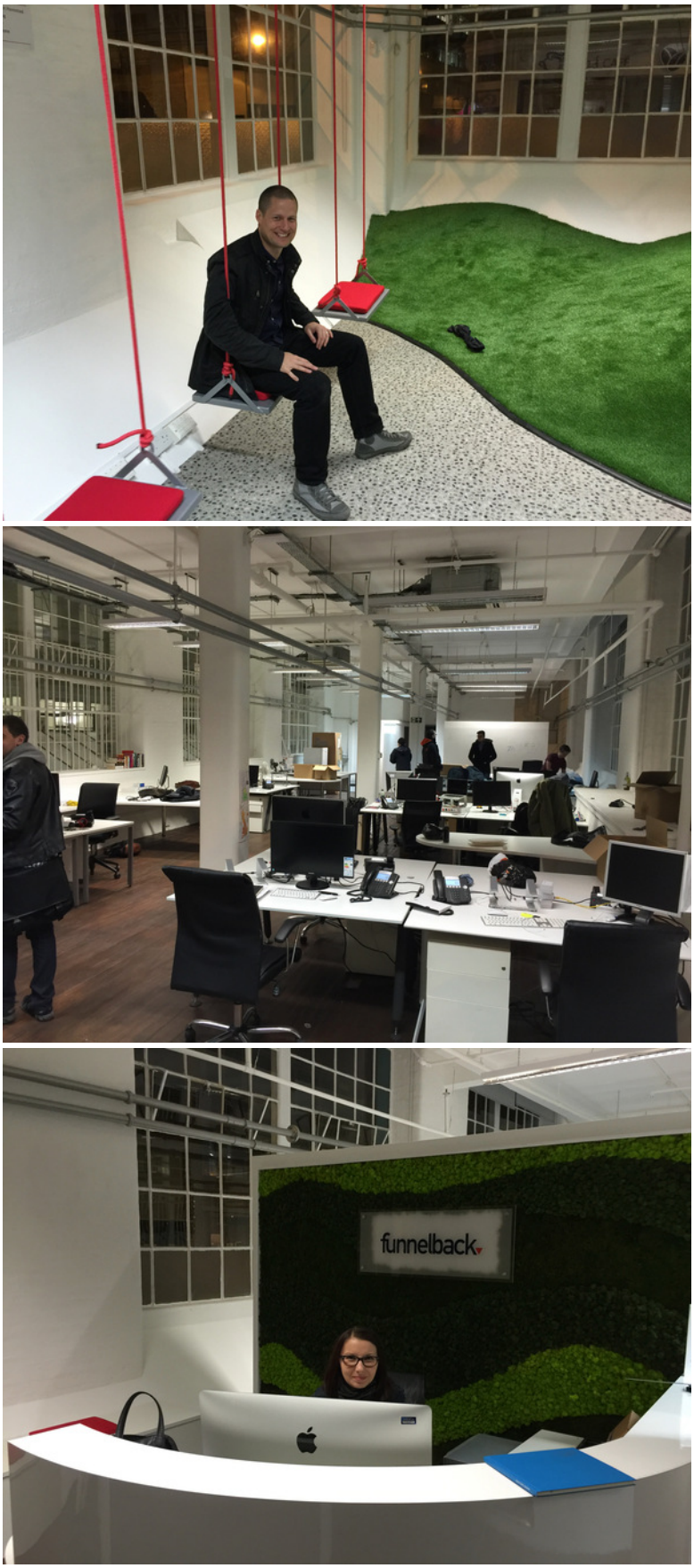
later told that he was the Dean of Westminster, but I'm not sure that was true.

Alwyn Davis and I were both excited at having the opportunity to present at such a distinguished venue. Westminster Abbey was in fact a Squiz client.

2010: Matt Taylor and Darryl Hannah staffing the Squiz stall at the Information Online conference held at the Kensington Olympia, London. In previous years that conference had been far larger.

2010: Matt Taylor presented at a Search Meetup, held in a pub in Borough Market, just near London Bridge. Alwyn Davis also attended one of these: "I went to a search engine meetup and it was my first time at something like that. Lots of smart people talking about stuff that was going over my head."

7.7 2015: The New FBUK Office



2015: Top: Gordon gets into the swing of working in the new office, on the ground floor of Zetland House. Bottom: The wall behind reception is covered with living moss.

2010: Naturally I always stayed in top hotels with impeccable facilities. Stuart Beil and I once shared a room to save money, but his pillow-throwing caused me to snore. ☺

7.8 Funnelback UK people

In the period after Squiz bought Funnelback in 2009, it was mainly Squiz staff who did Funnelback sales and support in the UK. Darryl Hannah, Steve Morgan, and particularly Phil Widdop did sales, and various Squiz implementors worked on Funnelback installations with remote support from Canberra. Eventually Funnelback UK (FBUK) was formed as a joint venture between Funnelback Pty Ltd and Squiz UK.

Alwyn Davis was Funnelback's first UK employee. He had left Funnelback and headed for the UK in 2009, where his partner Susan found employment as a dietitian in Birmingham. When Funnelback realised he was having difficulty finding a job, we asked if he would like to work for Funnelback in the UK. I "interviewed" him in a pub opposite Liverpool St Station and he was signed up soon afterward. I remember taking him to dinner some time later. At his suggestion we went to Brick Lane, which is a Bangladeshi community not far from the Squiz offices, heavily supplied with ultra-competitive restaurants:

– "Sir, you must eat at my restaurant. My mother is the cook and she has won many awards for her cooking."

– "No, no, no, sir, you must eat just here, where poppadoms are free and a glass of beer is included with every meal."

One of the work issues I discussed with Alwyn was his near-impossible workload:

> ... that first year is a bit of a blur because I had to do all the pre-sales meetings and demos, then actually implement all the projects we signed and finally support them all afterwards.

Alwyn returned to Australia in 2010 and continued to work for Funnelback in Canberra.

Peter Levan first encountered P@NOPTIC while working at the Australian Institute of Criminology (AIC):

> AIC became a customer in late 2004. We'd put out a tender to replace our existing search engine, which was a cobbled together `ht://Dig` installation. We purchased one of the final versions of P@NOPTIC (5.4 or 5.6 I think). And I recall there were two search servers, an internal one

which crawled everything (intranet, internet and the JV Barry Library catalogue; and one sitting in the DMZ which just served queries for the public searches and library catalogue (indexes were rsynced from the internal server if I recall correctly).

When Funnelback was released they upgraded. I'm not sure if they ever had 6.0 but they definitely upgraded to 7.0 at some point. Kat Ng and Ben Pottier did most of the work for AIC at the time.

How Peter came to start his career at Funnelback (12 years and counting) is rather interesting:

I joined Funnelback in late 2009. We were in London as my partner Sally [Greenaway] had been admitted to a very exclusive course at the Royal College of Music (RCM, which I think are now a Funnelback customer). I was looking for work there, getting really useless leads from a recruiting agency, when I remembered that Squiz had an office in London (I had no idea Squiz had bought Funnelback, or that Funnelback UK was even a thing). I knew Steve Smith from Squiz Canberra and just sent a resume into Squiz.

It was quite a coincidence that Funnelback UK had been set up just a few months earlier. Anyway, I received a reply from Squiz who had noted I had experience with Funnelback and they were wondering if I might be interested in a role (which didn't actually exist yet) with Funnelback UK. I was relieved to get an offer to work on Funnelback, even though it took a couple of months to create a job for me to do.

When I joined Funnelback UK it consisted of just two people - Phil Widdop and Alwyn Davis. Those two worked really hard - it was an unforgiving environment with so much work to do, and Alwyn especially had to wear so many hats. He set up the original UK SaaS, was presales/sales support, implementer, technical consultant and support person all in one. And back then we had very little communication options with AU. (It was quite some time before Steve Barnes set up an SSH tunnel back to Australia where you could try to contact someone via `Jabber`. Though it was complete luck if you managed to get anyone because the online statuses didn't work due to some LDAP issue and the time difference meant you had very little overlap. Despite that I did manage to have some great conversations with Steve Barnes and Kat Ng.)

Peter's partner Sally Greenaway[7] was one of five people from around the world admitted into the Royal College of Music's Masters course in film and screen composition. And when she flew to London to audition she was told that they had already decided to admit her! On return to Australia in 2010, Peter and Sally imported a large collection of traditional keyboard instruments.

In 2006, I was invited to give a keynote at the European Conference on Information Retrieval and was put up in Beit Hall at Imperial College. Within a hundred metres or so of my window were the Royal Albert Hall (out of shot to left), an imposing apartment building (left) and the Royal College of Music (right).

[7] `https://www.sallygreenaway.com.au/`

Peter Levan recalls his experiences in the UK

The UK was very different to Australia – the clientele was much less government and much more corporate. I also found that I didn't really fit in with the UK's pub culture – there were Friday afternoon standups that sometimes went until the early hours of the next morning while they emptied the beer cabinet (shower cubicle stacked with beer).

There were not that many actual projects while I was in the UK. They had one customer (the UK Electoral Commission) which Alwyn Davis had already set up. I implemented a search for Digital UK which was the first integration with Squiz's MySource Mini.

I also recall spending a bit of time with Justin Cormack (Squiz UK's head of technology at the time) trying to reverse engineer Funnelback OEM[a] – no one knew how it worked and it was basically undocumented. So we were trying things and hunting around on the Matrix server for logs etc. We figured it out in the end but as you know it wasn't a useful offering and never got much use.

Aside from these there were lots and lots of demos. In fact so many that I had to get Lisa Caswell to step in because it was stopping me doing paid work. I recall one of these was for a company that had terrible search. A lot of this was due to the fact that some idiot in marketing had decided that they needed a special character in their company name, and all occurrences of this in the text included the offending character. When it was indexed this character was dropped, so we had to write a filter to fix that up and the search started magically working after that.

The only other thing of note I can remember is having to go on a sales-fishing expedition to the University of Staffordshire (a long-time FB AU customer). One of the UK sales people had discovered they were a customer and wanted to upsell. I remember turning up there and the conversation went something like: "Oh yeah we use Funnelback. I'm not sure where that is though – we don't log in to it because it just works."

About half an hour later they'd managed to track down how to log in to the system and it turned out to be Funnelback 6.0 (and not touched since it was set up). This was a testament to the reliability of Funnelback, but also showed that the sales person hadn't done any research because we left the university without any interest in an upsell and a new-found discovery that they were entitled to annual support hours... (A postscript: Ben Pottier has since managed to upgrade them from Funnelback 6.0 - they are currently running 15.24.)

[a] See Page 207

Ben Pottier transferred from Funnelback Canberra to Funnelback UK in 2010 after joining in 2007. (See Page 130.) Still working in Shoreditch at the time of writing (October 2021), he is very close to Brett Matson's record for length of Funnelback service.

2011: Ben Pottier, back in Canberra for the Funnelback User Conference.

Ben Pottier on the unsuccessul Tesco project

We had a wonderful opportunity to provide a search solution for Tesco, which eventually led to you creating the beautiful solution of alternative metadata values with exception selectors. I clearly remember the review meeting with the Tesco head of IT, and explicitly the moment when he saw how this fabulous solution perfectly solved their problem of almost-identical data sets with minor variations. His face lit up with a big grin, and I knew we'd won it technically. I also learned from that one that no matter how good your technical game is, if you don't play the game at the C-level you won't win those. (We hadn't done that and the project was cancelled.) ☹

As noted elsewhere, I was a frequent visitor to Europe, often funded by organisations other than my employer. When I passed through London, I habitually made an opportunity to meet up with Funnelback UK technical people outside of the office, to thank them for good work and to discuss any issues they might have had. I first took Ben, Tash and Reyn to dinner at the Barbican. Later they preferred to meet me at their home, initially at a tiny house in Tooting Bec and later in a house and garden an hour or so out of London near Caterham.

Ben summarises his career at Funnelback:

> Overall, Funnelback has been my entire career. When I went to university we were explicitly told that the day and age of working for a single company your whole life was over and we would leap from company to company like locusts, a few years at a time. I don't regret sticking around at all, I have made many friends for life and learned more than I could have ever hoped for. From Stuart and Brett, to Gordon and yourself, I feel privileged to have worked with and known you all.

Next technical person to join FBUK was **Daniel Inniss** who was a fun-loving character of Jamaican heritage. He worked as a Funnelback technical consultant for five years, and had an abiding love for all ways of eating bacon. (Think bacon icecream!) He had a postcard of a mushroom on his pinboard, with an accurate label claiming him to be a Fun Guy.

2011: Daniel Inniss, technical implementer in the FBUK office.

Another technical recruit was Richard Morgan who had been a [discerning and very technical] client at the Victoria and Albert Museum.

2013: Richard Morgan, who joined FBUK from the Victoria and Albert Museum.

Phil Widdop joined FBUK as a sales person in September 2009, and became EMEA[8] sales manager in 2014, finally leaving the company in 2017. Phil says:

> It's easy to realise that Funnelback punched hugely above its weight. Being selected at Oxford University, Santander bank, NHS digital (among many others) on what was essentially a shoestring budget was only possible due to the product and the team we built ourselves.
>
> The Australian origins of Funneback always resonated well with the market (not too many other Australian SaaS companies around at the time) but looking back now, the main factors of the Funnelback growth in the EMEA were:
>
> - We built off advocacy in the UK education sector by delivering relevant applications to their needs
> - We sold search like an agency would sell a website; this was uber powerful as we pitched digital people.
> - We delivered what we said we would.
>
> Having since built sales teams for unicorns and companies that have taken on over $100Ms of Capital funding, I'm sure that, had Funnelback been given further investment in its product & marketing, Funnelback would more than hold its own with the likes of Coveo or Algolia.

2010: Hannah Cooper and Phil Widdop at the Conference Centre at Westminster Abbey.

[8]Europe, Middle East, and Africa

Gordon Grace in Brighton, after a meeting with Brighton Council. *Photo: Phil Widdop*

FBUK at a team offsite – clay pigeon shooting!: Daniele Pestilli, Ross Negus, Rokas Zalalis, Sam Arnold, Gordon Grace, Sonia Piton, Will Noble. *Photo: Phil Widdop*

Matt Taylor joined FBUK in June 2010 as Sales Manager and became General Manager a year later. In his two years as General Manager, FBUK head count grew from 3 to 13.

> My time in FB started in the summer of 2010. I had the opportunity to join IBM, but decided on a role that was a step up for me but also a slight gamble as the company was a very small start-up in London; the Australian search engine technology company Funnelback.
>
> I joined FBUK as Sales Manager and took on all new business sales and existing account management. Within 9 months I was promoted to General Manager to handle all aspects of the growth of the business. Even in that first brief period I could tell the company was positioned for success. We already had the likes of Dyson and Skype as customers (for our web search), but we were also able to grow and position ourselves as a really strong player in the market for Enterprise Search.

I remember handling the entire sales process for two incredible wins where we beat off the likes (I'm sure) of Endeca and Google, at Macmillan Cancer and University College London, the latter of which was the largest deal in FBUK history.

The UK team were great, and we were well supported by the larger UK arm of Squiz for things like marketing, design and finance. The head of UK tech, in particular, was a constant support in all aspects, as was the head of Sales at Squiz who helped me win many deals for FB. The Oz contingent, led by David Hawking and Brett Matson, were also a massive support. The way we pulled together as a small company and took on some much larger organisations and won, gave us all a sense of pride.

I had many trips around the UK, and the rest of world, with David Hawking, not only selling our solution, but also pitching up at trade fairs and watching David wow the audience. David was great at what he did, but also a fantastic host. I still fondly remember my Canberra walking tour with David trying to get me through my jet-lag. ☺

We also took the UK higher ed market by storm and must have had a large population of the red-brick and other Universities as customers – it was great to see. I don't remember many failures in my 3 years at FB, other than not being given the opportunity to continue my role in a more remote and flexible way; much to my incredible disappointment.

2011: Matt Taylor (General Manager, Funnelback, UK) on a jet-lagged visit to Canberra, relaxing in James Turrell's "Within Without" skyspace at the Australian National Gallery.

Sonia Piton is nearing ten years service with FBUK, having joined in January 2012. She had been a Programme Coordinator at the Journalism, Film and Media Department at Birkbeck College, University of London. Her roles at FBUK have included: Project Manager/ Account Manager, Production Manager, Services Manager, Delivery Manager.

Sonia Piton recalls some interesting projects.

Working with the **British NHS** team was a good experience – really clever, switched on team, focused on replacing their NHS Choices (main site) search to a a hard deadline (2 weeks implementation).
KAUST (King Abdullah University of Science and Technology). This project in Saudi Arabia had two firsts:

- Ben had to build a layer of Django CMS for it.
- I had to put an order to Funnelback purchasing for an abaya. ☺

2012: Sonia Piton at the Squiz UK event held at the English Speaking Union headquarters, London.

2015: Gordon Grace, Sonia Piton, Phil Widdop, and Ben Pottier at Curry Leaf East in City Road.

7.9 2009–2012: Recruits to Funnelback HQ

In this section I put together the stories of some of the people who joined Funnelback's Australian operation after the Squiz takeover, in the period before we moved to Allara Street.

Gordon Grace came to Canberra in around 2007 as a graduate entrant to the Australian Public Service. He joined the Australian Government Information Management Office (AGIMO) and within 12 months was working on Funnelback projects including publications search, govsearch and agency search. Govsearch rolled in to the Australian Government Online Services Portal (AGOSP) which evolved into myGov, and Gordon was heavily involved these projects. Around that time, Google had begun personalising search results for logged-in users, and suffered strong criticism for lack of transparency in how results were ranked. The Funnelback-powered AGOSP went to the other extreme, being careful to explain to users how their search was being personalised, and to provide a clearly visible link to "plain vanilla" results.

With AGOSP seeming to be entering a Business-As-Usual (BAU) phase, Gordon left AGIMO and joined Funnelback in November 2010 as a consultant in the Implementation team under Alwyn Davis. He saw it as a great learning opportunity and found he needed a fast ramp-up of his Linux skills. He found that he had a skill in writing good bug reports and found himself communicating

with the R&D team by this means. At the same time his talent and experience in building polished user interfaces led to him building more and more demos and delivering more and more sales and presales presentations. He also ran many training courses for customers and found he enjoyed it.

After Alwyn Davis left, Gordon became Head of Products and Services (P&S). Around that time, Funnelback's sales team quadrupled, and Funnelback began being sold to more and more Squiz clients. As Head of P&S, Gordon found more of his time taken in onboarding and upskilling staff, and more of Funnelback's work going into consulting.

Gordon was ambitious and, in 2014, when Matt Taylor moved on from the role of General Manager, Funnelback UK (FBUK), Gordon agreed to move to the UK and succeed him. (See his photo on Page 171.) He says that the UK move was the hardest thing he and his family have done.

Gordon feels that he moved to the UK at a very fortuitous time – lots of sales were being made, particularly in the university sector. Many British universities had deployed Google Mini and Google Search Appliances and found them less impressive than Google web search. They were also less than impressed that Google provided no support after two years, apart from selling them a new appliance. Even that avenue was cut off when Google stopped selling appliances altogether. Phil Widdop, an FBUK salesperson of long standing, had been pursuing the UK universities for years, and started closing deal after deal. In the end FBUK had around 57 UK university clients including four of the five top-ranked UK universities: Oxford, Cambridge, University College London, and Edinburgh. They also signed up the King Abdullah University of Science and Technology in Saudi Arabia. Gordon says that his visit to Saudi Arabia with Sonia Piton was very memorable.

FBUK also signed NHS Choices and its equivalent (HSE) in Ireland, the Royal Society, and some banks including Santander. Under Gordon's leadership, the Funnelback partner network was extended, and the Squiz team in Szczecin, Poland were trained up, leading to deals with banks and universities in Poland. Gordon really enjoyed his visits to Poland and his interactions with Rafał Źróbecki and his team.

All this commercial success led to a growth in the FBUK team to a peak of 16. Soon FBUK moved out of the underground office shared with Squiz UK and into its own (nearly adjacent) premises. The decision to rent the extra office space wasn't made by Gordon. He relished the separate identity and the extra space but worried about the burden of inner London rent. Although there was a continuing stream of new customers, it became increasingly difficult to maintain profitability. Eventually, the Squiz CFO came over to the UK and announced that FBUK was going to merge back into Squiz UK.

2011: A famous building at University College London, one of the world's top-ranked universities and, along with Oxford, Cambridge and LSE, a Funnelback customer.

Gordon remembers working with Verint, a major Squiz partner, which specialised in supporting local government systems and making joint pitches with them to the City of San Francisco and City of Vancouver. In some cities Funnelback was used as the knowledge access engine behind the 3-1-1 city service directory.

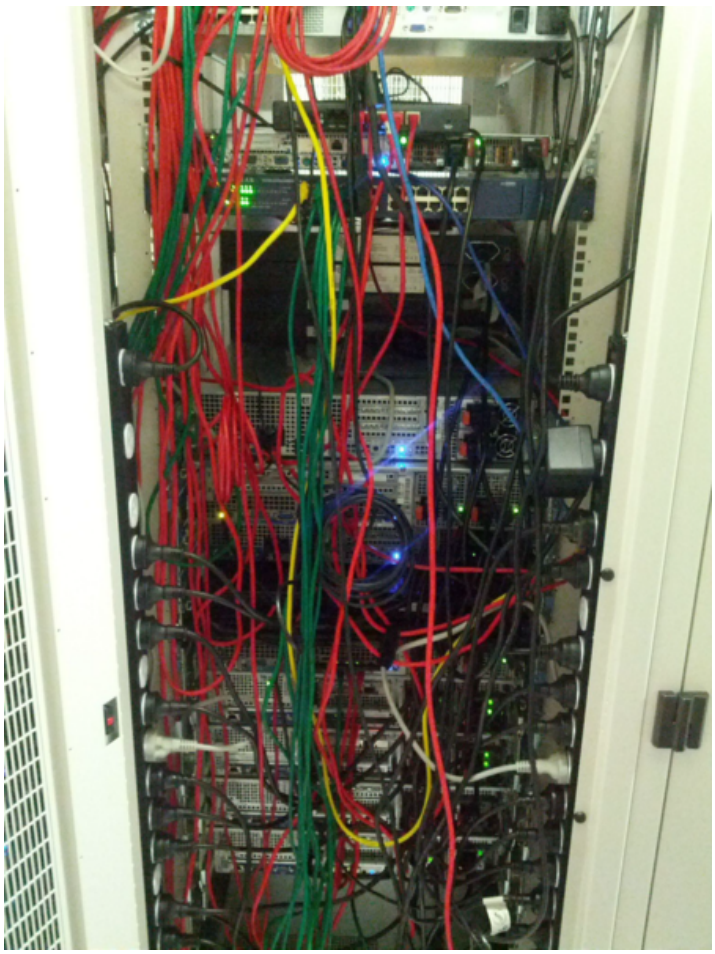
In late 2017 Gordon's family returned to Australia and he returned to the public service to join the Digital Transformation Agency.

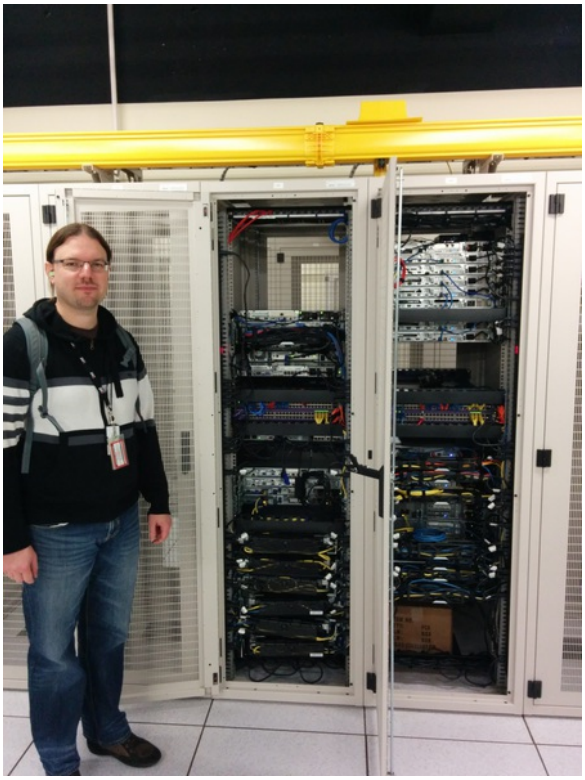
Nicolas Guillaumin (Nico) joined Funnelback in May 2010, having moved to Australia in January of the same year. His partner Céline Lenoble took a job as a teacher in Telopea Park (French-Australian) School and their visa gave him full working rights. He found that most jobs in Canberra were for government and that, as a non-citizen, he was excluded. One of the only interviews he got with a private company was with Francis Crimmins, who led the R&D team at Funnelback. Nico turned out to be a great appointment, and made vital contributions to Funnelback technology and hosting infrastructure.

> I recall that in my first or second week, I heard a scream at the reception desk, because Nalani Bakker saw a huge huntsman spider next to her. Brett put it in a glass and toured the office with the massive spider, showing it to everyone. I did not like that at all, but being my second week and Brett being the CEO I gritted my teeth when he presented the spider to me and tried to not panic! I don't know if he realized how uncomfortable he was making me!! There you go for my first exposure to spiders in Australia... [Dave: At least the spider wasn't a funnelweb or a redback!]
>
> From May 2010 to April 2013 I worked in the R&D team, under Matt Sheppard, who became the team lead shortly after I joined. Notable achievements in this period include the rewrite of the search UI from Perl to Java / Spring Framework, rewrite of the Trim connector from Perl to C#. I also worked on the design of the Continuous Updating system v2 but Luke Butters did the hard work of implementing it and ironing out the bugs.
>
> I worked on more minor things such as improving the filecopier system[9] for better performance, implementing some basic building blocks for multi-server support (e.g. separate crawler and admin, as well as publishing indexes between machines) but it was never used at a scale.
>
> I also spent quite a lot of time working on the continuous integration system (Jenkins) and the build, converting all projects from Ant to Maven for improved consistency and dependency management. I also re-built the automated browser testing system using Selenium 2/WebDriver, allowing us to quickly write automated tests for each new feature. I also wrote a regression testing system for the crawler so that we could see performance over time. It was an interesting project to "simulate the internet" by implementing my own web server and DNS server to trick the crawler into crawling from a single server. It resulted in various charts that were available in Jenkins and we could see for each change we made if it had any impacts on crawl performance.
>
> From April 2013 until May 2015 I became hosting manager after our previous hosting person Steve Barnes left and Shaw Xiao was left all alone to manage the SaaS infrastructure. He was quite junior and with no training in servers or networking. This was an intense time for me but very rewarding, both on the technical side and on the mentorship side with Shaw. Every change we made was applied to hundreds of servers and we saw an immediate impact in service reliability.
>
> The first few months were difficult because we had to continue putting out fires daily while trying to rebuild the infrastructure at the same time. Shaw was already using Puppet (a configuration management tool to manage and configure servers at scale) but not in the optimal way and only on 5% of the server configurations. I pushed a lot on Puppet – every time we had a change to make on a server from that point on we would do it via Puppet, implementing it as a generic configuration, rather than doing it ad-hoc on one server. It took more time initially, but after a while we started to be able to manage our servers more consistently, avoid specific configurations and improving uptime. In about a year we had turned things around – the platform was stable, and we were working on scaling it up and accelerating.
>
> From May 2015 to May 2016 I returned to the R&D team, and Simon Oxwell was hired to lead the hosting team. I think my work in this period was focused on revamping the Admin UI using a modern front-end stack: Angular JS. We experienced the pain of being early adopters, struggled a lot, and often had to dive into the AngularJS source code to troubleshoot issues. While doing that we also had to build REST APIs[10] for the system, since Angular JS would run as a client side application and make API calls to the back-end to administer Funnelback. I did write the initial API back-end but Luke Butters soon took over so that I could focus on the front-end.

[9] The utility for gathering files from a Windows fileshare.

[10] `https://en.wikipedia.org/wiki/Representational_state_transfer`





Above: Before the TransACT server tidy up. *Photo: Shaw Xiao.*
Below: Nico and Shaw beside their neat handiwork, after the relocation of Funnelback servers in the TransACT building. *Photos from Nicolas Guillaumin*

Highlights of Nicolas Guillaumin's time at Funnelback Australia

Implementation of a "one click deployment" of new servers. We used a tool called RunDeck to automate the provisioning of VMs, with Puppet taking over after the first boot to configure everything that was needed. Before this installing a new server took a couple of days and was very error prone as steps were always missed and mistakes discovered later.

Extended use of Puppet that allowed us to deploy changes to hundreds of servers in minutes, as well as collecting information on servers and the application across the whole SaaS very easily. A good example of this is that we implemented LDAP authentication on all the servers for FB staff, so that people could use their individual LDAP login rather than everyone sharing the admin password. With Puppet it was super easy to deploy this everywhere.

Revamp of the monitoring system (Nagios) closely connected with Puppet. Every time a new machine was provisioned, monitoring checks would be automatically added (OS-level, but also application level depending on the role: query processor, admin server, ...). We had automated checks for each collection, each profile, with smart detection of temporary failures to avoid false positives, etc. Before that the monitoring was so noisy that nobody paid attention to it.

Move of all the servers in TransACT from one computer room to another, without any search service interruption and minimum interruption of admin/crawling services. I'm really proud of that one as it's the first time Shaw and I did something like this and it went smoothly because we had prepared it a lot. I think it took us a few weeks to do it all. As part of that we also recycled 5 or 6 older servers and consolidated the hardware.

General stability and monitoring of the platform leading to a 4 nines[a] uptime (as measured externally by Pingdom), and close to 5 nines if I recall correctly. All of this work allowed Simon Oxwell to setup a US datacenter on the same model very quickly. Mentoring of Shaw was a great experience for me, and I believe for him. He put a lot of effort and we had a good connection. I recall we had weekly "mentorship meetings".

Redesign of the Admin Interface I spent a lot of time doing mockups and trying to figure out how to organize the new admin UI, with Steve Chan our designer at the time. I also got help from my partner Céline Lenoble (since she was transitioning into a user experience (UX) career at that time). She deserves the credit for coming up with the division between *Analyze* and *Optimize*, with 2 different colour schemes, etc. *Analyze* included the Search Analytics, Content Auditor, SEO Auditor and Accessibility Auditor. *Optimize* included Best Bets, Curator, Synonyms and Tuning. I did user testing, with internal staff, giving them tasks to do on the new Admin UI without showing it to them before to see if they could figure it out, noting where they got stuck, etc. It was a great way to see the pain points and improve the experience.

I recall when we were almost done with v1 and ready to release, it was given to Squiz for feedback. They came back with a complete revamp of the UI and colors, without trying to understand at all what we were trying to do UX wise. It was a huge disappointment and a negation of all the UI/UX work we had done and I recall becoming a bit emotional while talking about it with Brett in his office. In the end we managed a compromise where we would use the colors they wanted, but still keep the dual color scheme (blue / orange) for Analyze/Optimize. Still, I was quite pissed. ☹

[a] I.e. 99.99%

Natalie Grech worked at Funnelback from 2011 to 2020 in an evolving series of roles:

Funnelback contributed to my personal and professional growth in a way that I can hardly put into words. When I first joined in 2011, I had moved home to Canberra from Sydney. All I really wanted was a simple, part time role so I could focus on other pursuits.

In the end, I performed so many roles at Funnelback: office administrator, account manager, business development manager, content strategy, sales operations manager, marketing manager for ANZ.

Late afternoon in the Allara Street office: Greg Costin, Brett Matson, Natalie Grech. Greg took over Natalie's sales role when she went on maternity leave in 2015. ***Photo from Natalie.***

When I first started at Funnelback I was taken aback by the level of support I was offered by everyone one around me. I was in foreign territory with this technology and the sales environment, but I quickly began to realise I could ask questions, and not be dismissed or laughed at. I could walk into Dave Hawking's office and ask what must have sounded like the stupidest things. I could rely on Brett Matson for advice and mentorship and whiteboard diagrams. Narelle Bortolin pushed me hard for attention to detail and taught me to care about acting like a professional. Pete Levan painstakingly and patiently explained concepts I never would have thought I could understand. Matt Sheppard always made time to explain things to me when I felt frustrated and could not understand why we could, or could not provide a customer with the capability they demanded. Even before he took the role of head of sales, Ben Tilley encouraged me to back myself and was always so measured with his words even if there was a disappointing moment. And although it took me several years to appreciate, Stuart Beil taught me the value of diplomacy and networking when I always thought that kind of "stuff" was bullshit.

The spirit of camaraderie shared by the majority of people who worked at Funnelback was in no small part due to Brett Matson's leadership. And his leadership was supported in the earlier years by people like Dave Hawking, for whom he has great esteem and views as a mentor.

I remember a time when we met with an Australian government agency for a sales meeting and someone was in the room who wanted very much to be heard and to prove his point. Up until then, I had only ever seen Brett be congenial. The point being made was, "why would we choose Funnelback over (insert competitor name here)?". The person asking the question just wanted to own the room. And after about 20 minutes of being patient towards this argumentative character, very calmly, Brett looked him in the eye and said, "because not only are they exorbitantly expensive to the point of it being out of your league, quite frankly they don't have the technology we have." I thought, wow, this guy has some fight in him!

There are so many other people to acknowledge for their care and expertise and willingness - these are only a handful of examples.

The pivotal moment in my career at Funnelback was in late 2013/early 2014 when we lost several key players to our operations and management within what I think was about 3 months. I can even recall the dress I wore to work when I learned about the third loss. All of a sudden, I was the only sales person at Funnelback, with over 130 customers to account manage across

the country. Brett and I went to get a coffee, more for something to do than because either of us needed coffee, and we both reassured each other that everything was going to be alright. And it was. Between me, Brett, and eventually Ben, we visited every single customer we had, and reassured them we would continue to offer them the best product and service they had been promised in the first place. In that year, I don't believe we lost a single customer. And we had so many hilarious moments sitting in planes, waiting for planes, missing planes.

As described earlier,[11] **Peter Levan** initially joined Funnelback UK, but transferred to Funnelback HQ in 2010.

Peter Levan's projects back in Australia

There have been a few interesting projects over the years, though some are interesting for all the wrong reasons:

Digital Transformation Agency: This was an interesting project (courtesy of Gordon Grace, aka the ideas man) as it used Funnelback in quite a unique way. We wrote some really nice filters as part of this project that enabled us to do some very generic user-configuable content rule check. (So think of this as content auditor on steroids, where you can actually define the rules for what you are checking.) This enabled us to report on things like:

- Which sites/pages contain mixed content (search for http links within https pages)
- Which sites/pages have titles > 70 characters long
- Which sites/pages are applying rules from the Australian Government Design System
- Which sites/pages have non-English pages

I'm hoping that sometime soon this set of filters can be productised and placed behind a GUI that allows you to define the rules. The project was interesting because we didn't actually build a search index. The filtering produced a set of JSON files which were then loaded into Kibana for data visualisation. It was also a big data set, with over 20 million documents.

Transport Accident Commission (TAC) Victoria - accident data: We extended the work that was done for TAC to provide the search results as an interactive map. Funnelback indexed the (geospatial enabled) data and was configured to return search results in geojson format. This was then ingested into a Javascript library called Leaflet.js which handles the rendering of the map. It was a really elegant solution and enabled us to return very rich maps that included overlays such as LGA boundaries that could be added via KML files. Unfortunately, despite its elegance, the solution never went live.

Discover Tasmania This is quite a nice looking site and almost fully powered by Funnelback. Funnelback consumed the Tasmanian tourism database and was then used to generate almost the entire Discover Tasmania website.

[Dave: I remember implementing a commercially fair (i.e. random ☺) ranking option for Discover Tasmania. When successive visitors searched for accommodation in a particular area, they would see different orderings of the matching results.]

Some other memories:

- Turning up to the kick off meeting for the Sports Commission enterprise search and finding out that we had sold document level security (DLS) over 'their intranet' (unspecified). We found out that this intranet was SiteCore at the kickoff meeting and Brett and I exchanged looks, knowing that we didn't support this. Fortunately Matt Sheppard came to the rescue and created a solution.
- Working on the CSIRO intranet was painful because you'd spend at least 30 minutes on the phone to CSIRO support to get your password reset, as the password only lasted 30 days.
- Visiting a state government agency to look at their search quality. They were complaining that the search results were terrible and often the same when you entered different queries. It turned out they had done their own integration with Funnelback (XML) and their integration only ever passed the first search term to PADRE!
- In 2013, I worked on three very long-running enterprise projects: AMSA (Australian Maritime Safety Authority) upgrade, CASA (Civil Aviation Safety Agency) and ASC (Australian Sports Commission).

[11] See Page 172.

Prathima Chandra spent about seven and a half years with Funnelback after joining the team in 2011. She says:

> We had an amazing team, with a strong passion for technology and the company. I was amazed to see that the senior management team were indeed the creators of the software and the founders of the company. It makes a huge difference, as the passion for what we were trying to achieve as a company flows from the top.
>
> When Google was the search engine of the world, it was a great experience to sell Funnelback to replace Google Search Appliances (GSAs) and show to the clients how amazing Funnelback was. The fact that we had top government clients, universities etc. for more than a decade is a testament to the software. It was a great learning experience for me to learn about Search Engines and also about technology consulting. I have also learnt leadership from some exemplary leaders who lead by example and management skills.

A recent photo of Prathima Chandra. After Stuart Beil set up an office in Brisbane, Prathima moved there, working with him and Will Parkinson.

Will Parkinson also joined Funnelback in 2011. He was suggested to Stuart Beil by David Cousins, Bart Banda and Mark Henley who knew him from the web agency he worked at previously. He met Stuart and we got along well, even though he didn't know much about search at the time.

> We had two "interviews" (i.e. coffee and lunch). At the end of the lunch I remember him telling me, "If you get a haircut you've got the job."
>
> Working for Funnelback was a fantastic experience. To date, I cannot name a company I have worked for that I enjoyed more. I came into Funnelback knowing almost nothing about web or enterprise search, which meant quite a steep learning curve. Learning the ins and outs of the Funnelback product and company over the first 3 months in Brisbane and Canberra was a challenge but also immensely rewarding and enjoyable.
>
> All the staff I worked with were great. Pete Levan, Alwyn Davis, Gordon Grace, Matt Sheppard, Fran Crimmins, Nico Guillaumin, and Steve Barnes were always very helpful to me when I was starting out. The newer employees Gioan Tran and Phil Riethmuller also helped me out greatly at times where i was overloaded with work and needed a hand. Having the opportunity to work closely with Brett, Stuart and yourself was also very powerful for me. Working with

three people with deep experience in different fields allowed me to learn the many complexities on product development, running the Funnelback business and about technology and business in general.

In regards to customers, QBE Insurance, Queensland Health, The State Library of NSW, NT Government, Griffith University and the QLD Department of Agriculture and Fisheries have always stood out to me because they are always honest with their feedback, always willing to work in partnership with us to achieve a better search. We have also built a relationship with some of these businesses where we occasionally catch up socially.

Will Parkinson describes his experiences at QUT

I was on site with the Queensland University of Technology (QUT) for seven months on and off, developing their internal Enterprise search. We started the project in 2014 with a three week project and a defined scope. After four days working on site, the scope was changed multiple times and the project extended by weeks due to the inclusion of all content from student and staff systems into the QUT Funnelback search index.

This project was interesting and challenging. We didn't have a proper connector framework for Funnelback. At the time we were using Apache ManifoldCF which came with Sharepoint and Documentum connectors which were deemed their most important systems. These didn't work out of the box and I had to modify some of the code and submit it back to ManifoldCF for addition. I had to create a plugin for them to install into their Atlassian suite of products to produce lock strings to use for the Jira and Confluence DLS.

To enable DLS[a] for their Atlassian products in Funnelback, I had to alter some parts of the PADRE code to check for lock strings for Confluence and Jira. I ended up compiling my own `padre-sw-qut` to use as their query processor. Being asked to present updates to their large steering committee with less than an hour notice in some cases. I was at QUT for so long they added me to their staff birthday calendar and cleaning roster! I ended up forming a good relationship with the QUT search manager Simon Morecroft. I have met up with him for coffee multiple times since the project ended in late 2014.

[a] Document Level Security

Will Parkinson remembers travelling to Perth on a sales trip with Annie Pritchard to meet a government agency.

> The meeting invitation they sent us had us meeting one of their senior managers and two systems engineers. When we arrived we were confronted by around 15 technical staff armed with a sales presentation that they had already studied. The manager introduced me as a "senior engineer" from Funnelback (I definitely wasn't anything close to that, having only been at Funnelback for about a year) after which the group started to fire questions at me about how `PADRE` worked, the inner workings of the search algorithm, benchmarking tests against the query processing power of PADRE vs Solr, how the DLS mechanism worked, why we were using perl and C, which components I had personally built etc. They clearly were expecting yourself or a highly technical staff member. I think I managed to blag my way through it OK, even though the whole thing was terrifying, as they did become a customer about a year later.

In 2014, the prestigious ACM SIGIR conference was held on the Gold Coast. With Isabelle Moulinier of Thomson Reuters, I organised SIRIP – a Symposium on Information Retrieval in Practice[12] – as part of the overall conference. Although I was working for Microsoft by then, the program included presentations by Funnelback customers Maria Milosavljevic (Australian Crime Commission) and Kate Curr (State Library of NSW). Stuart Beil chaired a panel on commercialisation of IR research. Will Parkinson found the event an eye opener:

> Being able to talk to search experts working for large search companies was very enlightening. Being told first hand by some of the engineers at Baidu and Reuters on how they were going to develop their search over the next 10 years, with the introduction of AI, machine learning, etc. and then watching it since being rolled out was incredible. The way Baidu was going to utilise AI and image processing was both impressive and a little concerning.

[12] `https://sigir.org/sigir2014/finalsirip.php`

In 2013 we attended the Gartner Symposium on the Gold Coast. This was the single most bizarre work related event I had been to. A lot of C level and upper management types chatting and dealing during the day, but as soon as 5pm comes around, there were scantily clad models handing out drinks out everywhere, and a 1980's theme dress-up disco thing with terrible cover band. I found the frivolity rather strange.

Will ended up moving from Funnelback to Squiz in 2016 to work on some interesting Funnelback search projects there that integrated with Matrix and other enterprise systems.

I felt it was the right time for a change after you and Stuart had moved on.

2011: Narelle Bortolin looking concerned.

Narelle Bortolin was in charge of Funnelback finances for many years, before leaving in 2020. In his farewell speech, departing Chairman Stuart Beil highlighted the importance of her contributions:

Narelle plays a critical role that is often overlooked. She doesn't like the limelight, but she deserves the limelight. Without Narelle we would be in chaos. She has made my life and Brett's life so much easier. We have absolute confidence in what she does and in the integrity of her character. Narelle sees the most sensitive company data and to her credit discharges her role with maturity and confidentiality. She is extremely professional and competent and I have always appreciated the financial analysis she provides when I have requested it.

Narelle has taught me to love P&Ls and Balance Sheets, although my favourite will always be the cash flow statement and current financial position. I will miss singing a bit of Elvis or Flame Trees by Cold Chisel with you at those mid-year Christmas parties.

Narelle says:

Two of the best people I have ever worked with are Brett Matson and Stuart Beil. Their knowledge, professionalism and ability to bring people together for the success of the business is second to none. I would also like to call out Kiara Terwee as an inspirational, dedicated person who consistently brightened my days, as well as being the best Finance Assistant ever.

Mandhakini Iyer (Mimi) recounts her time at Funnelback:

I have vivid memories of my time at Funnelback. I moved to Canberra in 2011 after completing my master's degree in computer science in India. At the time, not only was I a temporary resident but I had little professional experience and the job market favoured permanent residents/citizens. Applying for jobs with no expectations was part of my daily life then. But I was pleasantly surprised to receive a call from Brett Matson, then Managing Director of Funnelback, who was interested in hiring me. After a couple of conversations, I landed my first job as a Technical Consultant in Canberra.

2011: Narelle looking much less concerned. ☺ Kiara Terwee, Narelle Bortolin, Brett Matson. *Photo from Narelle.*

2011: Mandhakini Iyer (Mimi) was a Funnelback Technical Consultant for eighteen months, then Service Delivery Manager for two years. *Photo supplied by Mimi.*

I had plenty of opportunities to learn and grow professionally at Funnelback. The opportunity to work on enterprise search projects across various industries in the early years of my career was enriching and set me up for future success. Two years since commencing my job at Funnelback, I was promoted to the role of Service Delivery Manager. This role exposed me to new areas such as product development, consulting services, hosting, managed services and service delivery, and the ability to work with cross-functional teams. In 2016, I had the opportunity to work across business units in Squiz Group and I took on a couple of complementary roles that leveraged my Funnelback product experience. These collective experiences gradually piqued my interest in Management Consulting, and I left Squiz Group in 2017 to pursue a career in Consulting.

Funnelback's culture was one of trust, and I was given the flexibility and autonomy to be my best. I also learnt a great deal about leadership from Brett Matson and Gordon Grace! When I botched up the client's intranet during my first technical implementation, Gordon and Brett showed me what it is to accept failure and learn from it. They always brought the best out of me, invested in my development, and acknowledged my effort. Dave Hawking, Tim Jones, Steve Barnes, Alwyn Davis, and Nico Guillaumin were always open to sharing their wealth of knowledge and patiently answered my repeated stream of questions about PADRE, templates, query logs, and implementations every time I approached them.

Tim Jones was a PhD student of mine, and a devotee of swing dancing. He tells his Funnelback story:

I have lots of fond memories of working at Funnelback – the *Funnelback coffee appreciation society* is definitely responsible for my current regular caffeine intake. [Dave: Tim took me to a PADRE coffee shop in Melbourne and I bought an appropriately labeled PADRE tin.] A few short quips:

- I remember Gordon offering to show me his tie collection for some reason – when I saw it, I commented that one of his was the same as one I had at home for dance performances – and he said something along the lines of, "See, that's the difference. I wear this stuff for work, and you wear it on the weekends."
- I did enjoy that you could view Matt's R&D team as a bad joke – An Englishman (me), an Irishman (Fran), and a Frenchman (Nico) walk into a bar to talk about software....
- I learnt an important career lesson when Luke (who was working with us on his research project) had a `php` question at lunch. I answered it in detail, before remembering that Nico used to be paid full time to do `php`. I said, "Hey, you know a lot about this stuff – you used to do it for a living", and he replied something along the lines of, "Why do you think I am staying quiet? I stopped working with `php` for a reason." `Php` (and other technologies I prefer not to work with) are no longer listed on my CV.
- Nico put together a cross-platform CD of Quake III (a first person shooter game and an old favourite of mine) so we could all play after work one Friday. Brett came in to the office at about 6pm to grab something, but was surprised to find us all playing. His reaction was, "Is that Quake III? Can I play too?", and he stayed for several hours.
- I don't know if you're including bug war stories – but there are two big ones – one was the PADRE function that was (correctly) creating a pointer on the heap that was valid before the function returned and corrupt afterwards. This only happened on the Mac, and it was because the header file wasn't included, so the function was assumed to return `int`. On a 32 bit architecture, this doesn't matter. On a 64 bit architecture, it "converts a pointer to an int" and truncates the value of the pointer. This was the catalyst for changing the code to prevent it generating any warnings on all (five?) of the platforms we compiled to, which I still think is one of the things I'm most proud of doing with software. At the beginning of that week I thought I knew C. At the end of that week, I definitely knew C (and had learnt a lot).
- There was another fun PADRE bug. The query processor was failing on very large indexes, which seemed to be corrupt. We used the index checker to find out what was wrong with the index, but it seemed to be buggy, because it said the index was fine – even after we improved it to be more stringent in the checks. Inspecting the query processor with the debugger revealed that there appeared to be HTML in the index, which was definitely wrong, because the index checker was happy. We eventually discovered that the query processor wasn't allocating enough memory to read the postings list and was continuing to process past the (truncated) postings. That happened to be into the memory that stored the snippets, hence the HTML. When we found this, we also found the comment in the source that allocated the memory for reading a postings list that said something like "this is a guess, but it's probably plenty."

Shaw Xiao initially joined Funnelback as a part-time system administrator in July 2011. He later transferred to full-time and his transitioned into more of a hosted dev-ops role. He worked almost nine years for Funnelback – a long time for his first professional job. He says:

> In 2011, I was a PhD student in the Computer Systems group at ANU. Although I learned many things from books and papers (but never enough!) about cloud computing, I was wondering what cloud computing was like in practice. When Steve Barnes interviewed me over the phone, I asked many questions about how Funnelback maintained high availability.
>
> Luke Butters joined about one month before me, also initally as a system administrator. We have maintained a good friendship since the beginning.
>
> I really got to know Nicolas Guillaumin when he transferred from the dev team to become leader of the hosting team. From his development background, Nico reshaped a system administration team into a hosting team and the role of system administrator into dev-ops engineer. I asked him to be my mentor, and he energetically carried out that role, while overhauling the way servers were administered.
>
> It was very enjoyable working with Nico on the project to migrate the TransACT (Canberra) data centre and to tidy the cabling. It was the noisiest (server and ventilation noise) and quietest (no chatting) working environment! (See the photos on Page 182.)
>
> I was originally ambitious, wanting to be a scientist in the company following on from David Hawking, but unfortunately that never happened. Dave is a great story teller. He was really good at talking about search technology and then seamlessly shifting to a travel story! I can visualise him sitting in front of The Desired Burger, having a double-shot macchiato, telling an interesting story. (BTW, Nico also picked up the habit of double-shot macchiatos.) I was really sad when the office moved away from Dickson and the legendary Desired Burger.
>
> Alwyn Davis has already talked about our multicultural Christmas lunch. It was a lot of fun, and I brought chicken feet just to make sure he would try them. ☺ Another fun thing I recall was the Frisbee game in Glebe Park near the Allara St office. I never thought a frisbee game could be competitive and tiring. Sadly I don't have a photo for that.

2002: I don't have a photo of Ultimate in Glebe Park either, but I have a photo of Frisbee throwing at the University of Melbourne: Nick Craswell throws to me while Alistair Moffat looks on. ***Photo: Hugh Williams(?)***

Luke Butters took over PADRE maintenance after I left. He has just recently left Squiz but recalls:

> Unfortunately we didn't really progress the pure IR side of PADRE, that is we didn't really do anything to make ranking better or query completion better.
>
> I did successfully re-implement *continuous* collections (now called *push*), so a user could send a document to Funnelback and that would be quickly indexed and made searchable. That was

mostly wrapping Java around PADRE and making a few changes to PADRE itself. I did add a poor man's parallel indexing to push, I think I had to slightly adjust the skipping of documents in `padre-iw` so I could merge the indexes together, that was fun and useful for a very large collection we had.

We didn't end up doing anything with the results of my investigations I did in my Computer Science Honours project. I think a super fast `vbyte` decompressor has been created so `p4delta` is no longer required, meaning `vbyte` is king. My project showed that using something like:

```
<doc-delta, occurrences, docOffset, docOffset .... >
```

was smaller and faster than the pair postings lists, but we didn't implement it.

I do remember Matt taking on one of your high-value roles:

Luke: "Hey Matt I need help with this."
Matt: "What is it?"
Luke: "It's Spring." [the java framework]
Matt: "No, it's summer!"
Luke: *rolls eyes*
Matt: "Dave's left. Someone has to keep making the puns around here."

In the seven years after my departure, Luke made many improvements to the PADRE suite of programs. To my delight, many of them were to reduce latency in various aspects of PADRE operation including query processing, index merging, continuous updating collections, and autocompletion. He also made various PADRE functions (like faceting and metadata operations) more convenient to use.

I find it rather rewarding that the code I started writing more than a quarter of a century ago has been actively used, maintained and improved to the present day.

7.10 Australia's Knowledge Gateway

Very early in 2011, we were contacted by Michael Gallagher and Kerrie Thornton of the Group of Eight (Go8) – a group representing Australia's most research intensive universities – about building a tool to make the Go8's research expertise more discoverable by potential research and industrial collaborators. The goal was very similar to that of DISR's Research Finder project[13] but the approach taken was quite different. The tool was to be called Australia's Knowledge Gateway (AKG).

The Australian government required Universities to report research statistics under the annual Higher Education Research Data Collection (HERDC) exercise. The separate Excellence in Research for Australia (ERA) exercise required them to provide full details of research outputs every three years. This was a massive data collection exercise for each university, resulting in a huge XML datafile.

It was clear that the ERA data formed the ideal basis for AKG. The universities needed to undertake neither additional data collection, nor build additional IT systems. Participating universities had only to make available their ERA data which for legal and financial reasons they were already required to collect. Fortunately, when an academic moved from one institution to another, the receiving institution tended to import the academic's previous publications.

Peter Levan set about creating an interface by which participating universities could nominate the URL at which their ERA XML file could be collected. He also defined a Funnelback primary collection for each university and an AKG meta-collection accessible for searches at `gateway.go8.edu.au`. The primary collections were scheduled to update at staggered weekly intervals. Individual universities were able to provide local expertise search, by searching only the relevant primary collection.

[13] See Page 90.

2009: Robed up for the PhD examination for Pavel Serdyukov at the University of Twente. From left: Djoerd Hiemstra (Pavel's supervisor), me, and Maarten de Rijke (University of Amsterdam). Both examiners are clutching professionally printed copies of Pavel's thesis. Although the outcome was virtually guaranteed, the public examination was high drama. Pavel stood in "the dock", accompanied by a paranymph[14] while being grilled by five black-robed inquisitors supervised by the Dean of the faculty and watched by dozens of spectators. Following the Dutch tradition, Pavel was later required to host a dinner for about 50 people during which staff and students from his department dressed up and poked fun at him and his work in a satirical play.

Scientific research into expertise finding conducted within TREC and elsewhere, including Pavel Serdyukov's University of Twente PhD thesis, indicated that the best way to rank experts was by counting the number of documents they had authored on the topic of interest. I built a search interface which, in its simplest mode of operation, received a research topic as a query – for example `angels dancing on the head of a pin` – found all the ERA titles and abstracts matching that query, extracted their authors, then sorted the authors in order of decreasing count of matching publications. The results page also provided a ranked list of universities based on author counts.

The results were not perfect. They depended on the precise wording of the query. They didn't take into account cases where an academic had moved on from an area in which they had previously been a prolific publisher. They were biased against junior researchers who hadn't had time to build up a strong publication record. If the search was for a very general topic such as *chemistry* the search was slow because of the massive number of matching publications.

Some effort was made to evaluate the quality of results, remembering that the goal was to enable external partners and students to find a short list of universities and/or researchers to contact for potential collaborations or PhD opportunities. Despite the deficiencies mentioned above, AKG seemed to support this goal quite well.

Go8 funded the development and operation of AKG for a couple of years, but they hoped to achieve participation (and cost sharing) from all of the Australian universities. Funnelback managed to sign up a small number but it proved to be quite difficult. Curiously, the Pro Vice-Chancellor I found myself negotiating with about AKG at Murdoch University in Perth turned out to live over my back fence in Canberra! That's some commute!

[14] Look it up!

Eventually, AKG failed for the same reasons that Research Finder had failed years before. Publishers of research wanted potential collaborators and students to search only their stuff, but searchers either didn't want to be so constrained, or couldn't remember the AKG URL.

7.11 Melbourne University: Find An Expert

After CMIS expertise finding, DISR Research Finder, CSIRO People Finder, and Go8 Australia's Knowledge Gateway, Melbourne University's Find an Expert was the fifth expertise finder I was involved with. All of them were quite different from each other. At the time Funnelback became involved, Find an Expert already existed, having been created by Simon Porter, and featuring powerful ways of displaying research relationships through collaboration and co-authorship graphs. Funnelback's contribution was in improving the underlying search. At the time of writing, Find an Expert still operates on the University of Melbourne site, but has much less prominence than it used to and no longer seems to provide the linkage visualisations.[15] Simon Porter left the University of Melbourne in 2015 and joined Digital Science Ltd whose headquarters are in London.

7.12 Squiz New Zealand

Squiz NZ is a Squiz subsidiary, located in Tori St, Wellington, and run by husband and wife team Patrick Fitzgerald and Kathy Olsen. Patrick is a quiet, behind-the-scenes manager while Kathy is a larger-than-life extrovert, with a passionate belief in social welfare and progressive causes. Kathy insisted that staff and visitors to Squiz NZ travelled in Prius taxis, and accessibility was a major theme at the Squiz NZ Christmas events. At one such event which I attended, a blind woman provided assessments of the accessibility of attendees' web sites by demonstrating what it was like to access them through a screen reader. It was a very compelling demonstration. The events were a family affair with Kathy and Patrick's son and daughter also involved.

2009: Me presenting at the Squiz NZ User Group held on the top floor of the tallest building in Wellington, NZ. ***Photo: Stuart Beil***

[15] I suspect it no longer uses Funnelback, but is rather based on a broader expertise management product.

2009: Down to serious business at the Squiz NZ User Group. (Well it was just before Christmas!)

2009: Stuart Beil, me, and Brett Matson in Wellington, NZ for the Squiz NZ User Conference. ***Photo from Stuart Beil***

Funnelback recruited some customers (including the Reserve Bank of New Zealand) directly and worked with Squiz NZ on a number of joint projects including the Auckland University of Technology, where Norman Goh was in charge of their web presence and search service.

7.13 Squiz Poland

Squiz has an office in Szczecin, Poland (not too far from Berlin), in order to benefit from the lower cost of employing very talented developers, and to establish a foothold in continental Europe. I met Rafał Źróbecki in the London office and later joined Steve Morgan in Szczecin to provide Funnelback training to the developers. We had hopes of using the Szczecin office to recruit European enterprise search clients. There were no successes during my time at Funnelback, but I believe that there have been some since.

2010: Rafał Źróbecki, Head of Squiz Poland. Rafał drove Steve Morgan and me to dinner at a rustic Polish restaurant furnished with unfinished timber and animal hides. The menu consisted almost entirely of strong alcohol and various types of meat. I was horrified that Rafał might want to drive us to our hotel but as we finished, Rafał's wife and a friend drove up and took us safely home. The next day we booked a bus to the airport in Berlin and the driver apologetically asked us if we'd mind travelling in his new S class Mercedes instead of the bus. Brilliant!

Completely undoing the pleasure of a luxurious ride to the airport was our experience trying to travel to London on EasyJet. The indicator boards said that our flight had been delayed by 45 minutes, but the truth turned out to be that the Berlin crew had timed out and that EasyJet were contacting on-call reserve staff in Luton, England to operate the London-Berlin leg of our flight and then fly us Berlin-London.

After getting to the head of a long check-in queue I was told I would be charged €45 for my bag and told to go to the long cashier queue to pay it, then come back and rejoin the check-in queue. After two hours at the airport we heard a whispered announcement that we were to go to Gate 29, which was actually just a ticket window. There they handed us an information sheet informing us of our rights and enough cash to almost buy a small sausage and a small beer.

Finally, at about 1am we joined the mad scramble for uncomfortable seats, whereupon the captain told us he was confident that the ground staff would have kept us informed of the reason for the delays!

On arrival at Gatwick, we were delighted that trains for London Victoria were scheduled at that hour of the night, and much less delighted that most of them had been cancelled. To add to the general air of luxury, there were no taxis at London Victoria!

2010: The building housing the office of Squiz Poland in Szczecin, Poland.

By the time (4.30am) I checked in at my hotel, I got the last free room – the one with the really tall window facing east, whose blind was missing the bottom metre. Since the sun was already up, on one of the sunniest days I ever experienced in London, I grumpily went back to reception and demanded a better room. Someone had already checked out so the manager went and cleaned that room, and by 05.30, I was sound asleep.

7.14 Squiz Scotland

Squiz maintains an office in Edinburgh at the very salubrious address of:

```
Royal Mile, 57-59 High Street, Edinburgh EH1
```

The office is located above a kilt shop and is accessed via a narrow stairway and door. It seems very unprepossessing but once into the office it's very serviceable. There's a reasonably sized training room overlooking the Royal Mile and a generous office behind with kitchen facilities and a shower. There's even a flat upstairs but the landlord (the kilt shop owner) seemed to discourage its use.

In January 2010, Paul Wheatley announced the first Squiz Scotland/Funnelback client – the East Ayrshire Council. After Paul's departure Guy Outram took over the reins in the Scottish office. I have fond memories of visits to Scotland, giving talks in the Royal Mile office and at Strathclyde University, and presenting Funnelback at SkyScanner, Glasgow University, Edinburgh University,

and the University of Dundee. The University of Dundee meeting occurred within an hour or two of Stuart and I landing in Edinburgh after flying in from Canberra!

2011: Guy Outram in the office on the Royal Mile.

After I left Funnelback, Guy inadvertently assisted in creating the illusion that I was a world-famous celebrity. I was holidaying in Scotland with my wife and my two sisters and, as we descended the stairs into Waverley Station, a voice among the heaving crowd called out, "Professor Hawking!" My sisters were duly impressed and it was great to catch up with him again, if only very briefly.

2010: Darryl Hannah (Squiz salesman not glamour actress!), Steve Morgan, and Stuart Beil near the Squiz Scotland office on the Royal Mile in Edinburgh.

2010: Steve and Stuart on the Royal Mile in slightly less clement weather.

2010: Stuart attempting to recruit graphic design students we met on the way down from Arthurs Seat in Edinburgh.

7.15 Funnelback in SquizVegas

Stuart Beil moved to Brisbane in December 2010. It had always been his plan to move back to Brisbane to send his sons to the same high school he had attended. It also was beneficial for Funnelback in that we already had a number of Queensland based clients including QUT and the Bank of Queensland. Importantly, the Queensland Government was also going back out to tender for the whole of Queensland Government search service and his presence in Brisbane helped us secure an extension to this strategic contract.

As Funnelback was growing so rapidly with clients in almost every capital city and overseas, it meant that Stuart was frequently travelling and it became less important which city he was based in.

Stuart recalls being in Canberra in January 2011 when he received a call that he needed to return home immediately to sandbag his house as the Brisbane floods started to rage out of control. Luckily his house did not flood.

The Brisbane Funnelback and Squiz offices were based on Montague Road, West End. The area had a certain vibrancy as it was transforming from an industrial precinct to a trendy inner city precinct with apartments, weekend markets and coffee shops. Stuart was fortunate enough to have a car park at the rear of the office that was accessed via a narrow laneway. He always felt uneasy leaving at night and one night his fears were confirmed as dozens of heavily armed police officers swarmed into the building across the other side of the laneway to arrest members of an outlaw motorcycle gang.

To handle the support requirements for Queensland clients and to help Stuart with pre-sales, Funnelback decided to recruit a number of technical positions for the Brisbane office. Prathima Chandra was happy to move to Brisbane from Canberra, as her husband had been posted to Brisbane. In April 2011, Will Parkinson joined the team.

2013: The Brisbane Squiz-Funnelback team shows its willingness to get down and dirty in the Tough Mudder. From left: Danny Peters, Bartek Banda, Tim Sutherland, Stuart Beil, Danny's brother, Mark Henley. ***Photo supplied by Stuart Beil.***

Stuart enjoyed having Will and Prathima in the office and every morning they would have a coffee in the cafe next door[16] and catch up on matters of importance.

Stuart recalls always feeling welcomed and part of the bigger Squiz team in the Brisbane office. He said "It undoubtedly stemmed from the office culture and personal relationships that had been established and it even went so far as Funnelback and Squiz entering a combined team in the Tough Mudder 16 km mud and obstacle course that tested our teamwork as well as our mental and physical stamina."

[16] Veneziano were not only a cafe, they were also wholesale coffee roasters!

7.16 R&D in the early Squiz years

Many R&D projects have already been covered in this chapter. Here are a couple of other projects that I was involved in.

7.16.1 *Query Auto Completion

In 2004, a Google employee named Kevin Gibbs developed a system for suggesting URL completions. When he showed his colleagues, one asked, "why don't you do that for search queries?" He did, and Marissa Mayer suggested that the feature should be called *Google Suggest*. However, it wasn't until 2006 that Yoëlle Maarek's team in Google Israel released the feature on `google.com`.

It may be that others had done it before, but when Funnelback built its version of query auto-completion (QAC), we made the feature quite general. Candidate completions could be drawn from any list of data pairs, not just from a query log. You could load the suggestions from a staff list, a product catalogue, or a list of famous quotations. You could include misspellings and synonyms. Suggestions were chosen using a formula which took into account both the weight of the suggestion and the difference in length between the prefix so far typed and the candidate suggestion. Even if the Barrier Reef is incredibly important, it's a big stretch to suggest `great barrier reef marine park authority` when a visitor types the letter `g`.

In addition to the standard mode of operation in which a list of most probable query completions/suggestions is shown, Funnelback's QAC mechanism also allowed suggestions to be snippets of HTML. This meant that photos, contact details, and biographical information could be shown for people, and photos, prices and catalogue information for products. Suggestions could also invoke JavaScript actions. This allowed current stock prices to be shown when the suggestion was a company or stock code, or stock levels when the suggestion was a product.

Finally, the list of suggestions could be segmented into categories with headings. Example segments at a university might be *Courses*, *Staff*, and *Research Papers*.

I really enjoyed building this feature with Nicolas Guillaumin – he designed and built the Java and JavaScript user interface and I built the C back end. We put considerable effort into reducing the latency as much as possible. A suggestion system which responds (including network latency) more slowly than human typing, is much less useful because by the time suggestions appear, extra characters have already been typed, likely rendering inappropriate some of the suggestions which appear.

My wife Kathy Griffiths is an experimental psychologist. She and I designed an experiment[17] to determine whether QAC enabled searchers to find what they wanted faster than using conventional search. It did. (The test was to find contact details for an ANU staff member.)

I also set up a demonstration of *faceted QAC* based on the stocklist of a hypothetical wine shop. Facets included wine type, winery, country of origin and price. Type `a` and the system makes 'a' suggestions from all of the facets. Select `Australia` and type `b` and the suggestions no longer include countries. Select `Brown Brothers` and type `c` and the suggestions no longer include either countries or wineries. With very few keystrokes and very few clicks, one can navigate to a wanted bottle.

BankWest were an enthusiastic adopter of QAC. Their homepage featured a highly prominent search box with QAC enabled. Among the candidate suggestions were individual web pages and even telephone numbers. For example, if a visitor types `c`, suggestions might include the landing page for credit cards, AND the telephone number to ring for credit card applications.

One serious problem with QAC is the reduction in user interaction data available for analysis and search tuning. A high proportion of visitors (I think it was above 60% at BankWest) who start typing into the search box never end up submitting a query, let alone clicking on search results.

[17] `https://david-hawking.net/pubs/hawking_griffiths_qac.pdf`

7.16.2 Geospatial Search

Some customers and prospects, such as Geoscience Australia (GA) and the Australian Broadcasting Corporation (ABC), requested the ability to impose a geographical constraint on queries. For example, ABC news stories were tagged with latitude and longitude coordinates, and the ABC wanted to allow people to search for news stories within their own area.

It was always possible to represent regional information within PADRE metadata (E.g. `postcode 3747` or `Cumberland Shire`) but at some point I implemented the ability to restrict search results based on latitude and longitude stored in adjacent metadata fields. You could restrict results to those within a specified distance of a specified origin, or to those falling within a specified rectangle. You could also rank results based on their distance from an origin.

A question arising was how to deal with search results which were not tagged with lat-long information. Do you just ignore them, or do you assume that they are at some default distance? In PADRE the behaviour was configurable.

Various organisations requested the ability to define regions using a region boundary – a long list of lat-long vertices. I thought about it, but didn't work on it. That functionality seemed to impinge on the space occupied by Geographic Information Systems (GIS) which provided capabilities way beyond the scope of Funnelback. I reasoned that the market addressible by such a capability within Funnelback would be very small.

7.16.3 DLS: Document Level Security

Will Parkinson's mention of PADRE lock strings reminded me that at some stage I designed and implemented the PADRE mechanism by which document level security was enforced. My memory is a little vague about details, but the basic idea was straight-forward. A metadata field was used to store a list of *lock strings*, such as `admin`, `hr`, or `Top-Secret`, obtained by the PADRE indexer from the gatherer or from *sitemap.xml*. The Funnelback UI was responsible for transmitting *key strings* to the PADRE query processor. After an initial set of search results was generated, the keys of the user were matched against the locks of each result document. If there was no match, the document was removed from the result set.

- There were defences against bugs in gathering or indexing which might result in a restricted document being indexed with no locks, and thus inappropriately included in results.
- In highly secure environments, each document in the result set returned by the PADRE query processor could be checked again for access immediately prior to display to the user, by masquerading as that user.
- In less secure environments where *translucent security* was desired, inaccessible results were included in the results set without summaries and without access to the actual document.

Chapter 8

Squizzleback

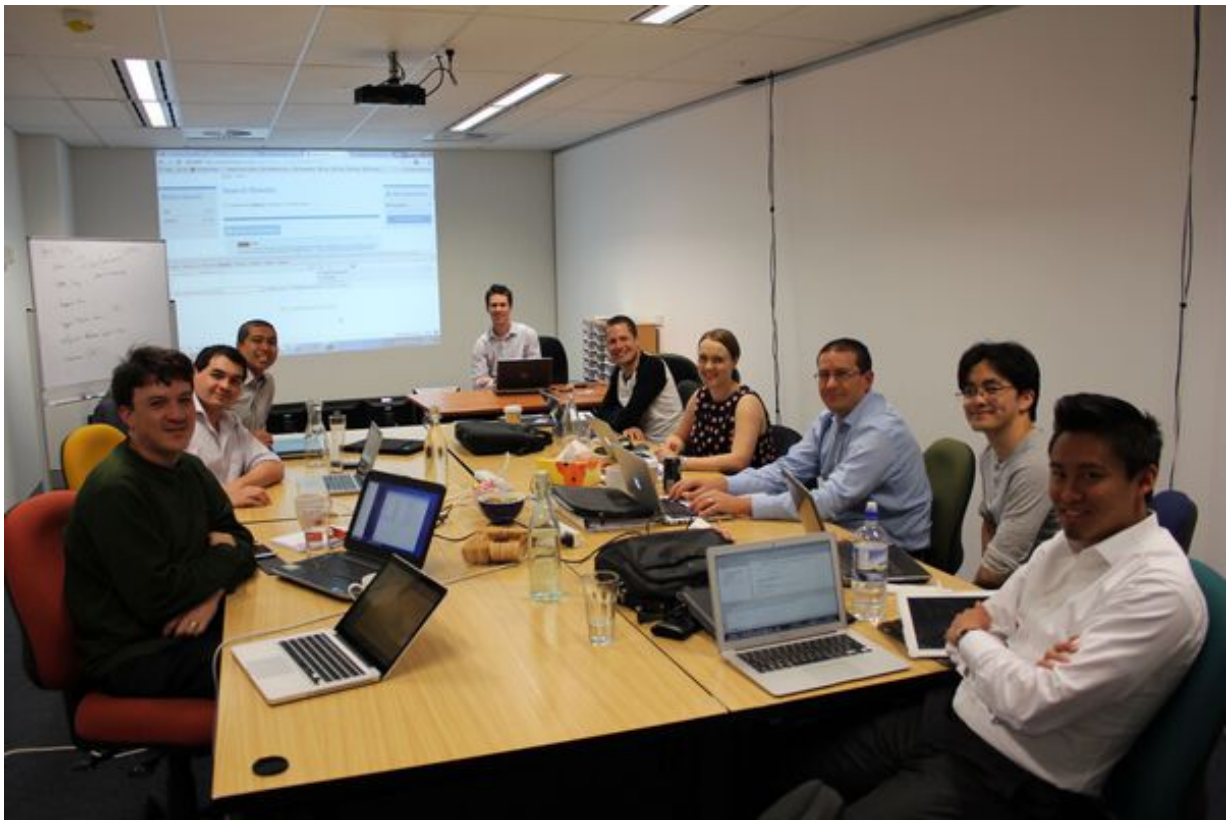
During the 2000s Squiz had built up a successful office in Canberra with Melanie Rooney as its manager. They were located in Deakin in office space whose lease expired in late 2012. JP and Steve Barker decided that it would make sense for them to co-locate with Funnelback in new shared premises. Accordingly, a search for office space ensued.

One possibility was to lease the upper-floor office suite adjacent to the Funnelback office in Challis St, Dickson, which had remained unlet for the entire time we'd been in Dickson. Squiz people who lived south of Lake Burley Griffin didn't like that idea. It was also felt that more salubrious options could be found. Various offices in buildings along Northbourne Avenue were inspected, but finally an office on the corner of Allara Street and Constitution Avenue was leased.

2012: Searching for office space in a prominent building on Northbourne Avenue. Melanie Rooney (Squiz Canberra Manager), and Brett Matson (Funnelback CEO).

Although I regarded Melanie Rooney (who I'd previously worked with at ANU) as a friend, and harboured no animosity to Squiz or its Canberra people, I thought that this was not a positive step for Funnelback. I didn't like the idea that Funnelback global HQ was being treated as equivalent to a relatively minor part of Squiz. I feared that it might be the first step in an escalating loss of Funnelback identity.

2013. An all-staff gathering in the Allara St, Canberra City training room. Left to Right: Adrian Khoo, Lachlan Henderson, Eu-Wyne Yeap, Gioan Tran, Stuart Beil, Ben Tilley, Narelle Bortolin, Shaw Xiao, Will Parkinson, Brett Matson, Gordon Grace, Anthony Barnes, Nathan Klein, Francis Crimmins, Scott Kaden, Vern Tee, me, Nicolas Guillaumin, Alison Bryant, Riaan Bredenkamp, Phil Riethmuller, Natalie Grech, Luke Butters, John Haynes. *Photo: Funnelback*



P&S team in Allara St. Clockwise from left: Peter Levan, Will Parkinson, Gioan Tran, Phil Riethmuller, Gordon Grace, Alison Bryant, Riaan Bredenkamp, Eu-Wyne Yeap, Vern Tee. *Photo: Nicolas Guillaumin*

8.1 At Least it Wasn't Collingwood ☺

The Allara Street office was actually very pleasant. The open plan working spaces were spacious and there was enough sound-absorbing partitioning to make it quiet enough to work productively. The R&D group were located in a psychologically separate area, furthest from the only external door.[1]

The quality of our accommodation was a relief after my experience of the Collingwood offices of Squiz's Victorian branch. They comprised a huge open-plan area underneath a sound-reflective, corrugated-iron roof. Hung from the roof were some large speakers constantly playing music. You could use a web interface to switch it to whatever it was you preferred. Think Beethoven's Fifth, followed by Doris Day, then Highway to Hell.[2]

To add to the din, a ping pong table was adjacent to the open-plan desk area. When one of the sales team finished a difficult call he would grab someone and head for the ping pong table where he would vent his frustration with violent smashes and loud expostulations.

The Collingwood office with ping pong table. Although a partition wall has been installed between the table and the desks, all the surfaces seem still to be acoustically reflective. ***Photo: Unknown***

Although open plan offices are promoted for their ability to foster communication and collaboration, nearly everyone in Collingwood sported ultra-heavy-duty noise-cancelling headphones and lost themselves in their own tiny world.

Tim Jones of the Funnelback R&D team had gained approval to move to Melbourne to pursue his swing dancing passion. He found himself in this curious environment and struggled to cope with the noise and the isolation. Matt Sheppard tried hard to support Tim by scheduling daily video meetings. Amazingly, the calls were sometimes difficult to hear at the Canberra end due to the ping pong.

[1] I confess that I may have a biased opinion because I spent most of my time in a luxurious individual office!

[2] As a supporter of the Essendon AFL club since 1960, it is my solemn duty to pour scorn on anything relating to Collingwood. Old joke: **Q:** What do you call a Collingwood player in a suit and tie? **A:** The defendant.

Tim complained to Squiz management and was given the use of one of the very very few offices, although he had to vacate it for a period during the middle of the day while a colleague expressed milk. Other people, forced to remain in the din, resented the special conditions given to Tim and his social isolation was complete. He soon left.

Tim Jones explains how the Collingwood speakers worked.

Because the speakers wouldn't let one user interrupt another's music (you could only change the music if nothing was being played right now, which I think is why the playing went for such long periods – because people didn't want to give up the control). One strategy I tried was playing silence for an extended period. This would work until someone would conclude that the system was broken, and reset the controller. Eventually I did tell someone I was doing this, under the argument of, "well, you've had your choice for four hours already today, this is mine", but as you can imagine, that was fairly unpopular. I think one of the sales guys was of the view, "I don't mind what music is playing as long as there's music."

(This was maybe not an example of my best behaviour!)

BTW, I agree with you about the Allara Street office. It was *excellent*, even without having my own office.

8.2 Further Integration

Over the following months and years, Funnelback's ability to operate independently declined and functions like HR, finance, marketing, and hosting infrastructure began to be integrated with Squiz's. The arrangements under which Squiz sold Funnelback licences also changed. All perfectly reasonable for JP and Steve (the owners) to make these decisions, but the exciting feeling I'd had of being a leader within a start-up gradually evaporated.

In 2013, I presented JP with a business plan aimed at increasing Funnelback profit. I argued that a reduction in reliance on billable hours and a focus on licences and SaaS revenue would be more likely to generate profit – we should aim to reduce the amount of consulting in each opportunity and substantially lift the number of customers. We should also aim to reduce the amount of sales effort needed to close a deal.

My view was that Funnelback hadn't fully recognized the diversity of the market for web site and enterprise search. Marketing messages generic enough to address all market segments failed to resonate with any of them. Messages needed to be targeted to the pain points and opportunities of individual market segments, like universities, e-commerce sites, call centres, law enforcement, local councils, government agencies, and large corporates. On the technical side, we developed products and services for dozens of different segments, but never had the resources to fully nail them.

There was also a lack of alignment. On the one hand, R&D were sometimes creating techologies which were not marketed or sold. On the other, sales people were sometimes selling technologies we didn't yet have.

I proposed an alignment of R&D, sales and marketing on single segments. I thought that the best opportunity was the huge university market. There were thousands of degree-granting institutions in North America and hundreds in the UK. Based on our experience with quite a few universities already, we had the knowledge to powerfully address the common pain points of course-finding, expertise finding, searching publications, searching learning management systems, detecting local plagiarism, and so on. Between Funnelback and Squiz UK, we had had extensive experience in building course finders. I had personally been involved in building five different expertise finders, and had examined a PhD on the subject.

Ideally, we would sell hosted solutions, which would provide a healthy ongoing revenue stream, and the ability to quickly resolve any problems which might arise. Our ability to cost-effectively tune the effectiveness of university search installations and to demonstrate search which promoted key university goals were strong selling points.

I knew that relevant university staff from across institutions regularly got together in meetings

and conferences, such as the Institutional Web Management Workshop (IWMW) in the UK, the Council of Australian University Directors of IT (CAUDIT), and no doubt similar groupings in the US and Canada. With a carefully tailored product, there would be a valuable opportunity to present, exhibit, and sell to large gatherings of target institutions. Surely this would increase sales and reduce the cost per sale.

My proposal was reinforced by Matt Sheppard's MBA project which used Funnelback as a case study. He found that there were more than 4000 degree-granting institutions in North America.

I thought that if this proposal went ahead we could re-create something of the original start-up feeling in Funnelback. I looked forward to the R&D, to contributing to messaging, to presenting to university audiences, and to helping with sales.

Toward the end of 2013, it became clear that, as was his prerogative, JP didn't buy it.

8.3 R&D Activities

As was appropriate for a growing company with a significant customer base, the R&D team under Matt Sheppard set about gradually rationalising operations. A trend over time was to replace `perl` with `Java` and the Java-compatible `groovy`.[3]

We used Atlassian Jira and Confluence to respectively track bugs and as a repository of discussion and documentation.

A dramatic improvement was the introduction of Jenkins.[4] On a daily schedule, Jenkins would check all the components of the Funnelback system out of our `subversion` repository and attempt to build it all and run the library of tests, on all the supported platforms. Any bugs or component incompatibilities making their way into the repository were quickly noticed and repaired.

For a time, major Funnelback releases appeared annually, just before the User Conference, and they were numbered according to the year – Funnelback 10 came out in 2010. By the time I left there was a move toward continuous deployment – i.e. the most recent version of the software which had passed all tests would be rolled out onto Funnelback hosted servers, and post-update tests would look for any problems and potentially roll-back to the previous version. This was a better way to go, but needed a lot of development to make it work. I'm not sure how far this initiative was taken.

8.3.1 Funnelback OEM

There is a very substantial overhead in crawling a large web site. Even in Funnelback's incremental crawling mode, every page must be visited to see whether it has changed. For a slow-responding web site the time required to do a polite crawl can be very long. For example, in 2009, at the UK Electoral Commission we observed response times longer than 45 seconds.

The *SiteMap protocol* provides a solution but few organisations use it. In essence the `robots.txt` allows a web publisher to specify a tree of XML files listing all the URLs from the site with date of last update, and optional external metadata.

It seemed that there would be a significant value-add for Squiz customers if Funnelback could provide a mechanism which fetched the SiteMap files from a site and updated only the URLs in the collection which had changed or been added since the last fetch, and Squiz enhanced its CMS offerings to produce an accurate site map. Access control metadata could potentially be supplied via `sitemap.xml` for internal websites.

I built a demonstration capability `Funnelback OEM` into PADRE. Instead of specifying a directory of files to index, you told the PADRE indexer to index a URL (or perhaps a list of URLs). In this mode, PADRE fetched the `robots.txt` file and the `sitemap.xml` file(s), updated its view of the data, and re-indexed. You may wonder why I built this into PADRE rather than building a separate fetcher into the normal Funnelback framework. Because Squiz used their own search UI rather than

[3] `https://en.wikipedia.org/wiki/Apache_Groovy`

[4] `https://en.wikipedia.org/wiki/Jenkins_(software)`

ours, I think that there was consideration of integrating PADRE into a standard Squiz installation, providing search over just that site, and using Squiz rather than Funnelback interfaces.

From memory, this good idea ran into problems because the pages on a site generated by a Squiz CMS, were dynamically created each time they were visited. Squiz developers seemed to think that it was too difficult to generate the required `sitemap.xml`, because of the dynamic way in which content pages were generated.

My prototype demonstrated that the basic concept was feasible, but I took it no further, given the block on `sitemap.xml`. It seems from Peter Levan's story in the panel on Page 174 that there was an unsuccessful attempt to get it going – I wasn't involved.

8.3.2 Funnelback Recommender System

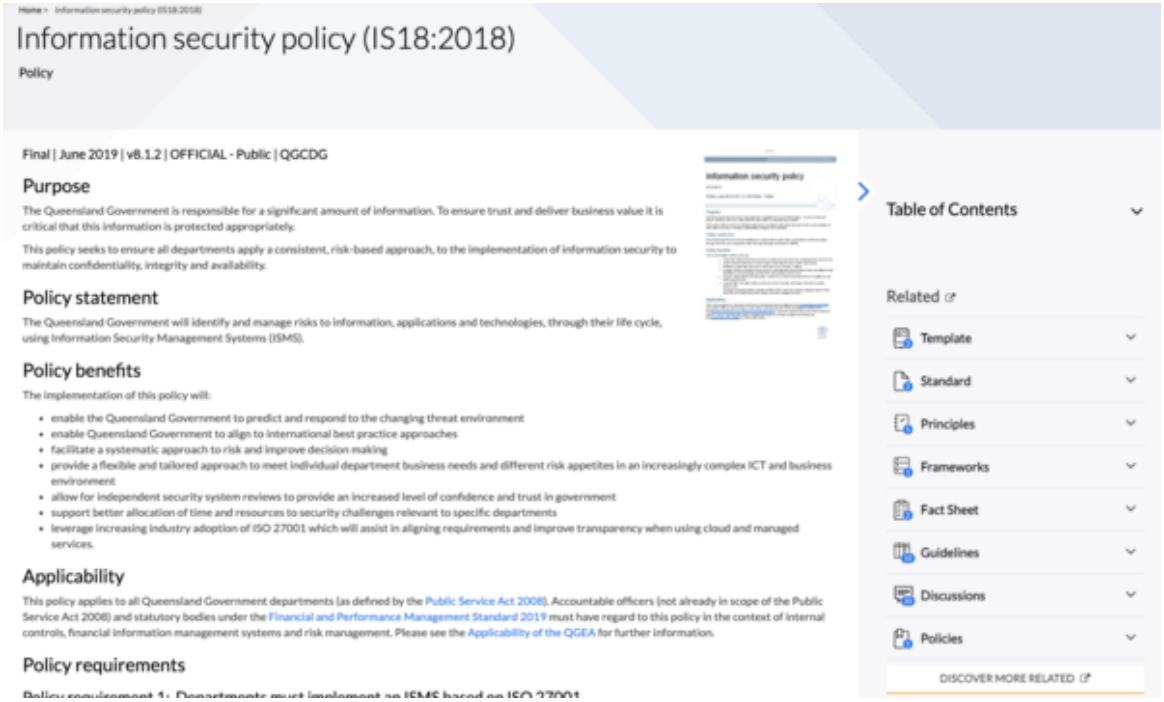
Francis Crimmins developed a recommender system to identify items related to a given document. The system combined information from multiple sources, such as co-clicks, related clicks, and related results. Most of the information came from mining query logs.

8.3.3 Funnelback Knowledge Graph

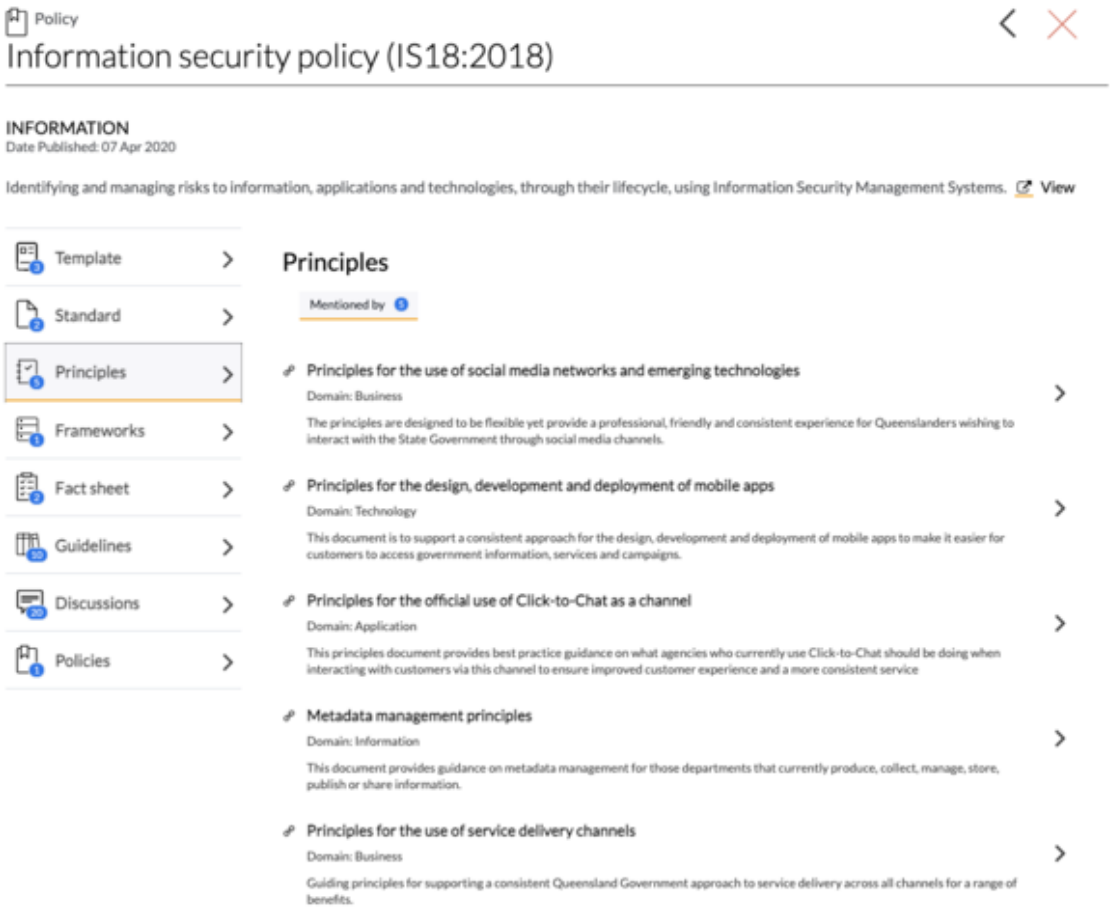
Brett Matson describes the enterprise knowledge graph which he was instrumental in building:

> The idea behind Funnelback Knowledge Graph was to solve the use case of finding all information related to a known item of interest, using a browse interface rather than querying. The item of interest could be a person, a policy, an event, a video, a guideline, a helpdesk ticket, an account, a sales prospect or anything else that could be represented digitally. Later phases of the Knowledge Graph roadmap included plans to use more complex graph queries to produce sophisticated analyses of the inner workings of an organisation. E.g. Has the organisational engagement of employee X dropped over the last 6 months? And how effective has HR been at engaging with remote teams?
>
> I had the initial Knowledge Graph idea, workshopped the UI with around 50 clients during an 18 month period, did a lot of testing and bug reporting, and wrote a 30 page training guide. The backend dev was done by Phaneendra Nakkala and the front end was done by Liliana Nowak (both in the R&D team).
>
> The MVP[5] was well received by many customers, including the Queensland Government Chief Information Office, who paid to have it implemented on their web site. Their use case was to help Government employees find all related information (such as information sessions, discussion boards, updates, guidelines etc.) to their various standards and policies (see images on the next page).
>
> Many universities, such as the London School of Economics, also expressed interest in using Knowledge Graph to promote their courses and academic expertise, but ultimately Squiz shifted focus to building its DXP,[6] and Knowledge Graph didn't get the sales, marketing or product development attention it needed to gain traction.

[5] Minimal Viable Product.
[6] Digital Experience Platform.



QGCIO: Viewing the Information Security Policy with Knowledge Graph shown on the right side



QGCIO: When you click on "DISCOVER MORE RELATED" (see image above), you can browse the graph directly...

8.3.4 Prediction Segmentation

Brett Matson explains Prediction Segmentation

Prediction Segmentation is an idea that came from Squiz CEO John-Paul Syriatowicz, who in turn got the idea from Marketo, a company Squiz was partnered with at the time. The idea was to use information derived from the search user's IP address to personalise search results. It worked by employing multiple datasets that enabled the search user's IP address to be mapped to a domain name, then to an organisation name, then to a Wikipedia article about that organisation, and finally the Wikipedia infobox fields for that organisation. This would allow search results to be customised depending on the end-user's organisation's industry, revenue, number of employees etc. For example, the Predictive Segmentation sales demo showed a search service for Squiz's own web site that would upweight case-study-related search results by geography or industry depending on the IP address of the search user and the industry/location of the organisation highlighted in each case study.

There were several challenges in using this in practice. End-users using a mobile connection were mapped to their telco provider, which obviously reveals nothing about the end-user's search preferences. Also, even when accurate organisation information for a search user was able to be retrieved, someone still needed to decide how the results should be personalised, which can be challenging to get right, and is damaging to search quality when done wrongly.

8.4 Funnelback USA

Ben Tilley transferred to Funnelback from Squiz after Annie Pritchard left in 2014. In 2016, he turned his attention to the potentially massive market in the USA, following a number of unsuccessful forays by others. In the beginning he started cold-calling US organisations during the night in Sydney. Eventually he moved to Seattle and, with the help of Nicolas Guillaumin (who also moved there) and Will Noble, was able to bring the US client roster to around 40 over a period of three years. Among them were the University of Southern California (USC), and Southwest Airlines. Ben left Funnelback in 2019 and Nico left to take up a job in Bern, Switzerland around the same time.

Ben takes up the story:

> Shortly after Stuart Beil left Funnelback, in Feb 2014 I think, I was offered his role and the opportunity to work alongside Brett and be in charge of sales and marketing globally, reporting to the board/owners of Funnelback (i.e. Squiz Pty Ltd/Steve Barker and JP Syriatowicz).
>
> Not long into my role I realized the future of Funnelback lay beyond AU and the UK. The sheer number of universities in the US meant it was a no brainer to go after that market. It had been tried before – Squiz sent someone over to build a Squiz office and they were not succeeding. Initially we tried to hire an ex Squiz staff member who was already based in the US. He tried to grow the business in the US but after two months he resigned, telling us that cold calling and growing something from scratch wasn't for him. We then tried to hire an external consultant to generate leads but that wasn't successful either.
>
> The only way this was ever going to work was if I did it myself, so I devised a plan to invest significant amounts of cold calling time while still running sales and marketing globally for FB.
>
> Three or four times a week I would go to bed at 6 or 7pm, set my alarm for 12am and wake up to start cold calling. I had a list I had painstakingly put together of directors and VP's at universities, split up into each timezone (so I could methodically move through each area and call people before they finished for the day or went to lunch) and I would cold call for five to eight hours straight.
>
> It worked! It took a lot of effort, but it worked. The primary goal of the calling was to generate initial interest and to organize the opportunity to meet in person. After a couple of months I had enough agreed meetings to justify me flying to the US. To make the most out of the time and cost invested in going to the US, I would map out my meetings. I would study a giant map of the US combined with flight and train options to work out the most efficient way to meet with everyone.

In the space of one to two weeks I would go to a new city almost every day and try to fit in two to three meetings in that day. It was a whirlwind of pitches, strange airport experiences, trains/buses, middle of nowhere towns, endless buffalo wings and fried pickles.

I made five trips to the US in a year, each one being at least one week and one or two of them being two weeks. I added in some conferences while doing these trips. This gave us the added benefit of meeting lots of potential prospects and building our presence.

One experience that will always stick out in my mind is the time I was left alone in an airport and asked to lock up. ☺

Ben Tilley closes up an airport

I flew into Laramie, Wyoming to meet with the University of Wyoming. I had caught an evening flight from Denver and I arrived at the airport at maybe 8pm. I called a taxi and was told there were only a handful in the town and the wait would be two to three hours. So I set up my laptop in the airport and prepared to catch up on a few things while waiting for the taxi. Slowly all the staff started to leave the airport, the TSA packed up, the stores closed, and eventually a lady came over to speak to me.

She said that everyone had left and asked whether it would be OK if she left too. She said I was more than welcome to wait inside the airport since it was pretty cold outside, and that once my taxi arrived I would need to turn all the lights off and lock all the doors before leaving.

So she took me around the airport, showed me how to check that the doors were locked then showed me where to turn everything off. She then left me alone in the airport by myself for a couple of hours. The airport wasn't large – I think it had three gates – but it was still quite an experience to be left alone there and then to get to lock it all up when I left.

Photo: **`https://flylaramie.com/`**

The response from everyone we met with in the US was incredibly positive – the Funnelback platform was clearly an improvement on everything else in the market that was readily available to higher education. Based on this positive response, I decided to move to the US to establish Funnelback North America.

I still retained my responsibilities as the global head of sales and marketing and I added to that responsibility for the entire North American business (i.e. MD/VP of North America).

I selected Seattle as the place to locate Funnelback due to its tech friendly business environment, lower cost of living and operating and proximity (from a timezone perspective) to Australia – which allowed us to benefit from the resources of the Funnelback technical team over there.

The first person to join the team in the US was Nicolas Guillaumin (Nico). He brought a unique perspective, having previously been in R&D, plus having been the leader of the SaaS team, and his involvement in the US was critical to its success.

Soon after, Will Noble joined from the FBUK office, and that completed the initially planned team. At first it was slow going. Although we'd built a lot of interest in the product we faced long

sales processes and a cooling of excitement once we started pushing people to sign on the dotted line.

Americans would also go quiet if they weren't ready to buy and wouldn't say anything. In contrast, in Australia and the UK you'd always be told if an opportunity wasn't going to progress.

It took some time to adjust to that, and to recognize that the initial strength of our business development pipeline was perhaps not as strong as we thought. However, we pushed on, continuing to grow the sales pipeline and starting to target verticals outside of Higher Education.

Building the team from scratch in the US gave us the chance to create our own mini culture within a larger company. It was a lot of fun and the majority of the time it felt like we were a start up. There were lots of table tennis competitions, pub lunches (an important corporate culture component imported from Australia), office brunches and all the fun things you do when you're a small business.

Every member of the team was critical to the success of Funnelback US. We got really lucky in that everyone was a great fit and brought a lot to the team.

Ben Tilley: Developing new product offerings (the key to success in the US)

Higher Education Product

After spending time analyzing why customers weren't buying Funnelback I came to the belief that what we offered was wrong for the US market. We were offering Funnelback in the way it was offered in Australia and the UK, which is to position it as an incredibly powerful search engine that can be configured in a myriad of ways – all tailored to the specific requirements of the customer. This type of solution was sold with large consulting projects and often with an unclear initial vision of what the final project would look like. [Dave: Not the sort of project I favoured!]

This was wrong for the American market. We had to complete against a significant number of additional software solutions all vying for the same budget pool and this meant that we competed against a lot of simple to deploy SaaS style platforms.

So Nico and I came up a product that did that. We called it the Higher Education product and we built the best example of a search for universities, with all the best uses of Funnelback features at the time, and sold it as a simple to deploy solution that allowed for custom branding but apart from that was an off the shelf product. [Dave: This sounds like what I pitched to JP in 2013.]

That description of it really simplifies what was built. Funnelback US owned the product development for this and due to that Nico was able to build a lot of stuff behind the scenes that enabled rapid deployment, easy upgrades, scalability etc.

Releasing this product was a game changer. To date I don't believe we ever signed a US based higher education customer to a normal Funnelback style project, all of them purchased this product that we developed in the US. Once we had this product rolled out to a few clients the rest started to fall like dominoes. In a short amount of time we signed up university after university and we started signing up big names.

2016: Ben Tilley sets up a huge Funnelback sign outside the WeWork office in Seattle. *Photo: Nicolas Guillaumin*

Ben Tilley: US partner program

Developing a partner program had always been a strategic goal for Funnelback. We were able to cut our teeth by partnering with Squiz, which gave us a sandpit to play around in and try to find what worked best, and then once we expanded to the US it was clear that to scale we had to partner.

To build the partner network we had to develop partner systems and we built a full partner portal which was powered by Funnelback.

This is one of my favorite accomplishments, because other people spent a year or two going around in circles trying to find the right third party platform to implement. I got fed up and worked directly with the technical team (I think Peter Levan led it) and we built a Funnelback powered portal in a couple of weeks, and it cost nothing.

It was also a huge success. We digitized our partner program and it elevated our appearance in the eyes of potential partners. We developed training programs (also led by Peter Levan I believe), a whole range of sales and marketing support and the commercial structure of the program.

We were successful in attracting several partners to Funnelback including another CMS developer, Terminalfour – who had an extensive Higher Education client base, and several digital agencies (mStoner and High Monkey). By the time I left, those partners plus Squiz US were starting to contribute significantly to the success of FBUS.

For a really small company FBUS managed to pull off some big wins. Here's some of the big names and some notes about them:

Southwest Airlines This one was one of Ben's favorites. FBUS was only included in the bid because the airline needed a certain number of vendors, and Funnelback was on the Gartner Magic Quadrant. Ben knew this would be a long shot and that FBUS would have to throw everything at it to win it. He flew down there to meet with them (based on the correct assumption that no other vendors would do that) and had the team put together an amazing demo, including really cool things like including flight data from an API so you could enter your flight number and get live information about the flight.

Lincoln Financial Group This was a big name financial institution, another one we didn't think we had a chance to win.

New York State (whole of government contract) This was solely a WCAG/Section 508 auditor deal but a great name to win!

State of Indiana (whole of government contract) Search and accessibility deal for the whole state of Indiana including agency (to reuse an FBAU term) search. This whole deal was done by Will Noble and it was massive.

Washington University in St Louis A very prestigious university. Top 20 I believe.[7]

Fordham University

University of Southern California

Nicolas Guillaumin's recollections of Funnelback USA

In June 2016 I moved to the US, following Ben Tilley to open and develop the US office. What a great opportunity! Céline and I wanted to get a bit closer to Europe and I also wanted a change of role, so it was good timing. For the first 6 months or so I still reported to Matt and did mostly R&D work (working on the new Accessibility Auditor) because there were no projects yet in the US. But soon we signed our first customers and I was in charge of implementing the projects, with support from the Australian team.

Initially it was just Ben and me. We had a small office in a WeWork co-working space, on 107 Spring Street in Seattle. That was tough for me because he was on the phone almost all day, and doing a lot of cold calling, always with the same script. It was hard to concentrate and deliver R&D work during that time. Working remotely with the R&D team was difficult due to the time difference. Still, the life in Seattle was nice, and the office atmosphere was great. It was full of startups, lots of interesting people to meet. There was a ping pong table in the basement, and beer on tap (a fact we kept reminding the rest of Funnelback staff in Australia!), regular events such as weekly happy hours and "one minute pitch" events for the startups.

[7] 18th in the US according to Times Higher Education. https://www.timeshighereducation.com/world-university-rankings/washington-university-st-louis

We almost felt like a startup ourselves because it was a small team and we had a lot of freedom. (At least that's my impression. I'm sure Ben had some pressure from above to produce results.) At the same time we had the backing of the whole organization so we did not have many of the struggles of younger/smaller companies.

In October 2016 we were joined by Will Noble from the UK office. We moved to a bigger four person office, in the same building. Soon after that the US election happened with Trump being elected. We went into a pub with our partners plus a US citizen John that we knew from WeWork to watch election night. For us it was quite appalling, but for John it was terrible – as soon as the result was known he left the pub crying.

Ben and Will were focusing on higher education initially, since we were strong at that in Australia. In parallel I started working on a *higher education package'*, a set of pre-configured collections and templates that we could quickly copy to implement for new clients. There was a typical web search, a course finder (with shopping cart, a directory search, social media search (YouTube/Twitter/Facebook channels of the university), events. We used a look and feel from a Photoshop mockup provided by Squiz. It looked good and modern and made for good demos for prospects. On the technical side it was implemented via `Stencils`, a set of templates and Groovy scripts and classes that were sitting on top of the product, but still managed as a library that could be installed on existing Funnelback installations. Initially it was developed by Gioan Tran in the Products and Services team, but I was able to take it up to the next level due to my software engineering background and work in the R&D team. I had a better chance of integrating it more closely with the product. You can still see screenshots of this on the Squiz web site: `https://www.squiz.net/products/funnelback/for-he`

We started winning small colleges and universities, in California if I recall. I did join my colleagues in pre-sales sometimes. I implemented almost all of the US projects, then later on we got help from the AU colleagues.

In April 2017 Paolo Sciarra joined us, as a *demand generation manager*. His sole role was to generate leads for Ben and Will, and he did a great job, with a mix of advertising, e-mailings, brand awareness campaigns. He even made Funnelback socks to send to clients, I still have my pair! He found a lot of prospects that Ben and Will would then contact to try to get a meeting. He left after a year I think – he wanted to move to New York and change careers.

I treasure my pair of Funnelback socks!

In July 2017 Jesse Swingle joined us as the US Marketing Manager. He was great too, did a great job at producing content (blog entries, and social media), working on the web site, on conferences/events, etc. I was quite critical of Marketing before (like a lot of IT people!) but seeing him work convinced me there's value in it, and that it made a difference – that's how good he was. ☺

In November 2017 Owen Pickford joined as a Products and Services consultant, to help me implement projects, as we were getting more and more clients. Mostly higher education, but also a few "random" ones like the International Risk Management Institute (IRMI), Agilent (a pure Content-Auditor deal, no search), a medical doctor association, and also Southwest Airlines! I think part of it was that we were in a few reports like Gartner thanks to Squiz so we started popping up on the radar of IT deciders.

Owen did not have a lot of technical experience and wanted to learn more technical skills. However he had good experience as a consultant and business analyst, and a great mindset as he had set up the Indian branch of his previous employer on his own, in India. It was another good experience for me to mentor him on the technical side (but not needed at all on the consulting side!)

2017: Paulo Sciarra (left) and Jesse Swingle in the third Funnelback USA office at Industrious.
Photo: Nicolas Guillaumin

When he joined we took another two person office in the same building on Spring Street. So we had one room with four people (Ben, Will, Jesse, Paolo) and one room with two people (me and Owen). However it felt a bit crowded and we started to tire a bit of the WeWork mindset and the small offices. Ben started looking elsewhere and found a good deal with another co-working company called *Industrious* which was more upscale (and expensive) than WeWork. It was at 2033 6th Ave, Suite 600, in Seattle downtown. Just next to us was the new Amazon buildings including the gigantic spheres.

The new office was great. We had more room; it was more cosy. The common area was way better than WeWork, higher quality furniture, less cramped. The view was nicer too. The staff were really tending to our needs, and they organized original events regularly. I recall once for Valentine's day some chocolate maker came and taught us how to make praline chocolates for our significant other. Another time it was a micro-brewery doing a beer tasting and explaining how they brew beer, etc. I really enjoyed the time here, – the best office experience I ever had!

2018: Amazon building (a.k.a Jeff's Balls) across the street from Funnelback USA's new office.
Photo: Nicolas Guillaumin

In March 2018 we were joined by Hannah Fuchigami as a Marketing Specialist. She was working under Jesse and executing his strategy. At the same time Jesse took over Marketing globally as there was no-one for Funnelback Australia, UK, etc. I think Squiz did some at some point but it didn't go well and/or the person that was supposed to be 50/50 between Squiz and Funnelback did not end up spending a lot of time on Funnelback. So it was decided that Jesse would take over, since he did such a great job in the US.

During that time Ben and Will were trying to break into the financial institutions market. It was tough, almost impossible to get a hold of anyone, all deciders / IT managers were super protected. But he did manage to get in touch with Lincoln Financial Group, a big investment and insurance institution. They contracted us to build their web search, and intranet search as well.

One thing we realized while doing the business analysis was that they needed "content level security" because their portal system on their intranet presented different content boxes depending on user permissions. They didn't want the crawler to crawl the "admin" view otherwise people would be able to see restricted content via results summaries. It took them a long time to explain their requirements, then it took us even longer to make them understand the challenge. In the end we found a workaround but I don't remember the specifics. The project was not finished when I left at the end of 2018. We also had to enlist Sonia Piton (from the UK office) to do the project management, because they were very process oriented. She came with us to Greensboro NC to do business analysis workshops. I recall that after the first day we went to the restaurant and ordered a double single-malt scotch for each of us! ☺

During that time we were also trying to build partnerships. We did get close to a company called *High Monkey* in Minneapolis. I think Ben met them at a higher education conference. They were doing content management projects with various clients including higher education institutions. We had a good connection with the two founders, Virgil on the technical side and Joel on the commercial side. Virgil was actually a star in the Sharepoint world and did a lot of conferences as a speaker, including on search topics. He was very excited by what Funnelback could do. They were also partners of Kentico, a CMS vendor and Virgil and his team built a Funnelback connector for it. I did visit them quite a bit to do workshops and training sessions with their teams, and they came to Seattle too.

I travelled regularly at that time to deliver training for US clients, since I was the only consultant in the US. I especially recall one trip in February 2018, which was not in the US but in Dublin, Ireland, to meet with another partner and train their staff. I can't recall the company name unfortunately. They had their own CMS and worked in higher education as well, including in the US, and wanted to build a packaged solution with CMS + Funnelback. I traveled to Dublin and met with them the first day, but then a snowstorm hit and I was stuck in my hotel room and them at home. I did the rest of the training (one week) on Skype or Zoom with them from my room, which completely defeated the point of having traveled all this way! I could have done the same from the US. ☺ Still it was not too bad and I got to visit the Guinness factory over the weekend, despite the fact that Céline who was supposed to join me from France was not able to make it due to airport closures.

In mid-2018, Ben moved to Chicago and opened an office there. It was also just a room in a WeWork, at 20 West Kinzie Street. It was not clear to me why the move was made. One reason was was that Chicago was a better airport hub than Seattle.

In mid-2018 I knew I wanted to leave, partly to go back to Europe and partly because I had been working for Funnelback for a long time, in different roles, and wanted to see something else. So we asked Owen if he wanted to take my role, and we looked for another consultant. David Mikulis joined us in the end of 2018, in Chicago.

I left in December 2018, we had a joint Christmas / Farewell party. It was quite emotional for me because it's been a good chunk of my life (8 years), over 2 different countries and 3 different roles. I did learn a lot and I think I brought a lot on the table as well, and I was really super proud of what I had achieved, especially in the US. We started from nothing and after a few years had an office with staff, partners and substantial yearly revenue.

2017: Funnelback USA Christmas party. *Photo supplied by Nicolas Guillaumin*

2018: Nico's farewell. At back from left: Hannah's partner, Hannah Fuchigami (Marketing Specialist), Jesse Swingle (Marketing Director), Nico, Ben Tilley, Céline Lenoble. Bottom row: Will Noble (sales, from the UK office), Owen Pickford (consultant/implementer). *Photo supplied by Nicolas Guillaumin*

2018: Nico's farewell. Nico with one of his presents, a photoshop of the Funnelback US team faces over the Metallica band, re-using the Metallica logo style. They knew I was a Metallica fan. ☺ On the right is David Mikulis, new consultant from Chicago. ***Photo supplied by Nicolas Guillaumin***

2018: Nico's farewell. Hannah Fuchigami transitions from virtual to actual reality and seems a bit shocked. ***Photo: Nicolas Guillaumin***

8.5 2013: Dave Leaves Funnelback

By 2013 I felt that Funnelback leadership was no longer driving Funnelback's direction. I felt that Funnelback had lost its independent identity and its start-up feel. I no longer felt like a co-driver / navigator of a team on a mission.

I think JP hoped that Funnelbackers might seamlessly transfer their loyalties to Squiz, but its culture and business model were different. Although Squiz has been an amazing success story over more than two decades:

1. It didn't share my goal of developing a scalable business focused on high quality search.
2. It didn't have an R&D focus.
3. It was founded, owned, and controlled by other than the Funnelback leaders.

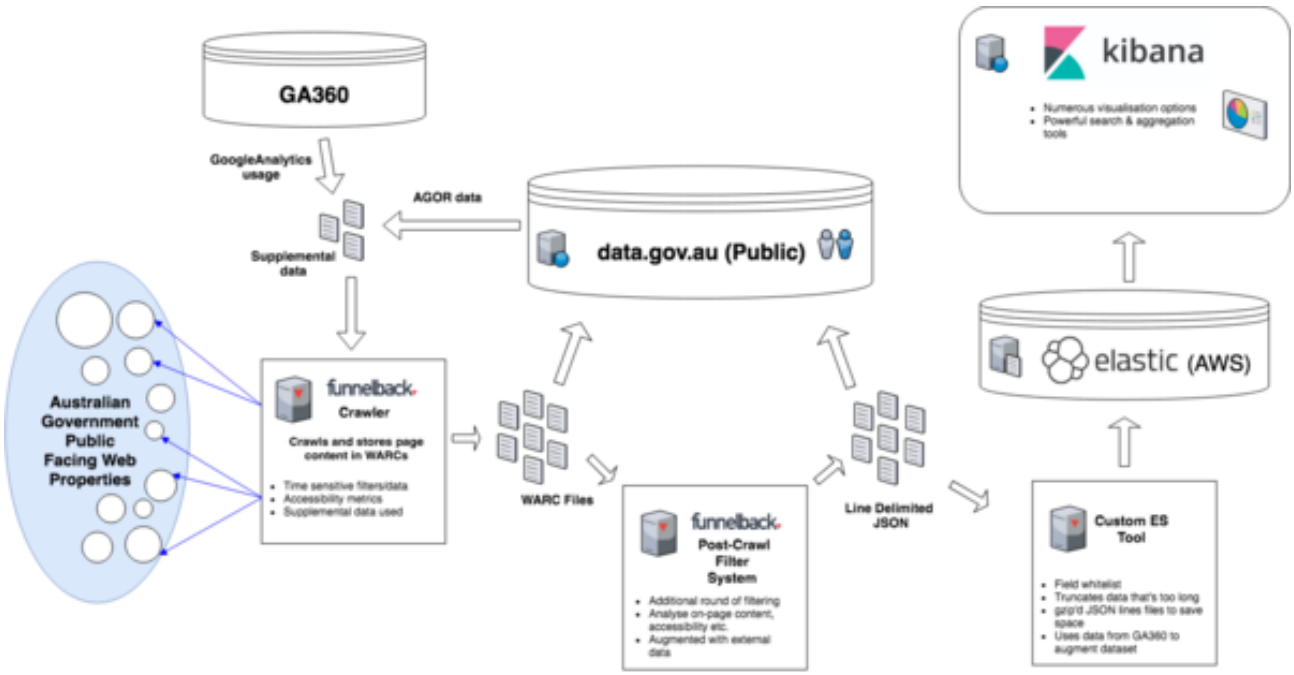
Coincidentally in 2013, my former ANU/CSIRO colleague Peter Bailey was at the time returning to Canberra, though continuing to work on Microsoft's Bing web search engine. He and his boss Nick Craswell (my former colleague in ANU, CSIRO and P@NOPTIC) asked if I would like to join Peter in Nick's science team. Another round of interviews ensued, and in November 2013, I left Funnelback to join Bing.

It was of course hard to leave behind a technology which I had worked so hard to create, and a company into whose success I had put heart and soul, but the time was right. The fact that Microsoft gave me a big increase in "compensation" provided a little sugar coating.

Stuart Beil, whose sales expertise had led to a huge proportion of Funnelback's total sales revenue for the preceding decade, also left around the same time. Of the original Funnelback Pty Ltd team, Brett Matson was the last person standing.[8]

2013: Q&A session at the Funnelback User Group meeting, held in Old Parliament House Canberra. Stuart Beil presiding. Gordon Grace, Matt Sheppard, me, and Brett Matson fielding questions. ***Photo from Stuart Beil***

[8] Like me, Matt Sheppard was a secondee to the original team, but he had a couple of gap years away from Funnelback.



c. 2019: The high-level architecture of the content auditing system created for the Australian Digital Transformation Agency.

8.6 After My Departure

Because I departed in late 2013, the Funnelback story after that must be told by others.

Nicolas Guillaumin and Ben Tilley have written in detail about Funnelback USA, which was created after my departure. See Page 213. Luke Butters has written about his subsequent work on PADRE. See Page 191.

I'm told that Australian Government projects to audit the compliance and accessibility of government content published via the web, continued. The diagram above illustrates the complexity of the most recent in a long series.

Brett Matson comments about the 2014 – 2020 period in Australia:

> I think the interesting point is that Funnelback eventually adopt the focused Higher Education strategy, but ultimately Squiz did too. In both companies, this included launching Higher Education specific products that were backed by sector-dedicated services, sales and marketing within a matrix structure. This strategy gave Funnelback the focus it needed to break into the challenging US market and achieve the highest sales momentum the company had ever seen (around 80 customers by the time I left). Within Funnelback, the next phase was going to be to repeat the strategy for additional sectors, but Squiz eventually changed focus away from solving sector problems to modernising the platform.

8.6.1 Jack

In a way my connection with Funnelback continues. When I retired from Microsoft in March 2018 I organised a retirement breakfast at the Margaret Whitlam Pavilion in the Canberra Arboretum. I invited work colleagues from my various past roles, reaching back to a Beechworth cherry orchard almost 50 years before. Brett Matson was of course there, and unknown to me, got into technical discussions with my son Jack Griffiths-Hawking who had recently returned from five years in the Philippines. Brett was sufficiently impressed with Jack's knowledge that he invited him to interview at Funnelback.

It's not nepotism I tell you! I had nothing to do with it.

Jack has worked for Funnelback (now Squiz) for more than three years, first as a Technical Implementer, then briefly as a member of the R&D team (contributing some code to support for WCAG

2.1), before being transferred to fill a gap in the Hosting team. He worked on a significant project for the Digital Transformation Agency. (See the diagram on Page 220.) He used the `pandas` Python library for data analysis and also worked on a natural language prototype for one of the big four banks, using the `spaCy` Python library developed by Matthew Honnibal and Ines Montani.

His current position is Hosting Engineer and he reports to Simon Oxwell. He makes extensive use of Python and `Rundeck`, an open source tool for automating administrative tasks on multiple machines. A common task is format-converting OpenSSL certificates.

He tells me that although Funnelback hosted servers are still located in TransACT in Canberra, they are being migrated to the Squiz cloud whose servers are located in Sydney, Melbourne, New York, Sacramento, and in the UK at Ash and Newbury. The UK sites are run by a company known as *The Bunker* which has repurposed an underground nuclear bunker in Newbury, Berkshire, and a former military radar station (Ash) near Sandwich in Kent.

Top: 2019? Clockwise from left: Kiara Terwee, Jack Griffiths-Hawking, Adrian Khoo, Greg Costin relaxing at the end of the day in the Allara St office, Canberra City. ***Photo: Funnelback***

Chapter 9

Significant Frustrations

Over the years, a number of projects have given rise to significant frustration. Forgive me for getting some of them off my chest. On the other hand please forgive me for being circumspect about some of them, and for leaving others out!

In the early years P@NOPTIC was included in a large consortium bid (involving Telstra among others) for a major Victorian government opportunity. A huge amount of work was involved. A motorbike roared across Melbourne to deposit our bid but arrived minutes after tenders closed!

A reseller submitted a joint bid with Funnelback (or was it P@NOPTIC) for a government project. When successful, they cut Funnelback out of the bid and used a competitor search engine to increase their profit.

In the very early days, we offered to provide a low-cost search to a government agency which had an extensive library of publications. Providing an effective search capability would have been a significant value-add.

> **Public servant:** Too much trouble. We'd need to go to tender.
> **Dave:** We could provide it for a year at a cost below the threshold for going to tender.
> **Public servant:** There is no such threshold. I'm required to go to tender whenever there is the possibility of cost savings for the government. I buy library books for the agency and I tender for them!

Really??? The cost of running a tender must surely exceed any possible savings on library books. For the tenderers, the cost of responding to a tender is likely to drive up the price!

Although Squiz acquired Funnelback in order to add effective search to its content management and web experience offerings, the Victorian branch of Squiz several times bid competitor search solutions to avoid having to share revenue with Funnelback.

I remember Brett Matson spending a huge amount of effort preparing a bid to be included on the approved IT supplier panel for a state/territory government. The bid wasn't accepted and when Brett asked why, the official in charge said he had binned it without reading because of some trivial issue on the front page. When Brett asked how that approach would deliver value to the government, the official said he didn't give a flying ____.

At a federal government agency Stuart Beil remembers that someone internally had not gone through the proper "competitive tender" process and just decided to buy a Google Search Appliance (blue box), reportedly for $1m. He was appalled and protested that Funnelback had not had the opportunity to be considered. His appeals fell on deaf ears. In the end we heard that the blue box became an "expensive paperweight in the agency offices."

There were a number of frustrations at the Department of Immigration. Peter Thew was working on a proof-of-concept project illustrating P@NOPTIC's ability to index and effectively search the TRIPS incoming and outgoing passenger database, when our project was abruptly terminated and IBM was brought in at a cost of hundreds of millions of dollars in response to the embarrassment of the agency having deported an Australian citizen Vivian Solon and unlawfully detained another (Cornelia Rau). (Our project was valued at a few thousand at most.)

Later, we were given to understand that Immigration would engage us on a project to search several important repositories. Because we were already a supplier to Immigration, we were told by our contact that we could be awarded a contract without a tender. (Our offering had been compared to those from other existing suppliers.) However, when it came time to sign the contracts, the procurement officer refused to sign, because our existing contract was far smaller than the new one. If we had been Microsoft or IBM, it would have been no problem because they already had large contracts. Ours was just too small. The project would have to go to tender. ... Unfortunately, it would take months to write the tender documents and run the process, and that would tip over into the next financial year, by which time the funding for the project would have evaporated! ☹

Another federal agency told us that they were very much in favour of our product and that the delegate would sign off on the contract "before she goes on leave this afternoon." What our contact at the agency didn't know was that a revised bid had been received from a multi-national competitor, and they were now the winners. Grrrr!

It was profoundly disappointing that government agencies, federal and state, were so consistently unsupportive of local companies. Government contracts, *of course awarded on the basis of a fair evaluation*, could have made a huge difference to the success and growth of Funnelback, but time and time again we found that government purchasing was biased toward large companies, almost all of them headquartered overseas.

Chapter 10

Why Didn't?

10.1 *Why Didn't We Become "Google"?

In the 1998 *Anatomy of a large-scale hypertextual search engine* paper, Brin and Page reported that their index contained 24 million web pages. At about the same time, the ACSys WAR project was distributing the not much smaller VLC2, a collection of 18.5 million web pages. VLC2 was comparable in size to the data indexed by many of the then current web search engines. By July 1999 PADRE was able to index (with shortcuts) VLC2 in about 7.5 hours on a single Pentium 3. However, at that stage our `pwget` crawler had no hope of crawling such a large number of pages.

With the benefit of hindsight, I find it interesting to try to answer the question, "How did Google rather than P@NOPTIC come to dominate web search?" Fortunately, I can do this from a position of comfortable retirement and with no regret over lost opportunities to own a 100-metre yacht.

Nick Craswell reports a chat he had with Larry Page at WWW7:

> I asked him about how he evaluated relevance, and he said "when something is better you know it's better." I went away to study Web search evaluation and he went a different direction.

Here are some factors that I think played into the "why not Google" question:

1. I started down the path that led to S@NITY/P@NOPTIC/Funnelback with absolutely no consideration of commercialisation. It wasn't until pressured by ACSys / CSIRO that I started thinking about it. Even then my goals were to earn enough to be able to keep on doing the research, and to find out about real-world search problems.
2. In ANU's Department of Computer Science at the time there was no culture of commercialisation. We had no commercialisation precedent to follow.
3. We started from the small and narrow, e.g. TREC and the academic discipline of IR, the problem of searching and managing intranets. I never took a big picture view. I.e. "Here's this rapidly expanding thing called the web — what would it take to index it and provide fast search service? How could one do a better job than Alta Vista, HotBot etc.?" (Nick and Paul were way ahead of me in realising the potential of the web and learning about web technologies — I'm grateful to them for dragging me along.)
4. TREC participation got us into the game and provided a wonderful leg up but later it distracted us from focusing on commercial success. All that time trying to persuade the academic Information Retrieval community that they should be interested in Web search! We built test collections instead of a world-beating web search engine. Although we drew the attention of our prospects to TREC results, they never influenced a single P@NOPTIC sign up.
5. We weren't adventurous enough to jump out and form a start-up. In around 2000, Darrell Williamson, CEO of the ACSys CRC commissioned John Fick and his company iFund to prepare the ground for a P@NOPTIC spin-off but none of us were really interested.

John Fick's assistant, a consultant, had a couple of good jokes about his occupation:

> Q: What's the definition of a consultant?
> A: Someone who borrows your watch in order to tell you the time.
>
> Q: What's the difference between a consultant and a shopping trolley?
> A: You can put more food and drink into a consultant but you get better direction out of a shopping trolley.

6. In Australia at the time, there was virtually no chance of getting investor funding to build a global search engine. Steve Kirkby, who later became a Funnelback Chair, told us that investors would want to see our "skin on the fence" – that we should be prepared to mortgage our houses. No, thank you! We were constrained by lack of investment both for software development and for hardware and network traffic. In contrast, the Google Anatomy paper thanks IBM, Intel and Sun for donating hardware, and also thanks their "funders".

   With our focus on intranets and enterprise search, it would probably have been difficult to get significant investment, even in the USA.

7. In the critical phase prior to 2000, we didn't have business people with the right skills. In contrast, I think Brin and Page had very solid advice from investors and from Stanford people who had experience with start-up companies.

8. Our technology was too limited, and advanced too slowly because of organic growth and too many distractions. With more rapid development, PADRE might have cut the mustard — it started with multi-server capabilities and eventually got to be able to index and search (with full capabilities) up to about 100 million pages on a single large server.

   Our Funnelback crawler would have needed a lot of investment to quickly scale up to crawl the whole web. Writing it in Java and using Java objects and garbage collection caused many problems. The model of doing complete crawls[1] for each update was also inadequate — it probably needed to be database driven and to do clever selective updating. It would also have needed to be given the capability to operate on a cluster of machines rather than a single server.[2]

   Nick Craswell's wonderful perl scripts got us up and running in very quick time but starting the perl interpreter was too slow and they wouldn't scale — we should have quickly built low latency versions capable of operating at scale.

9. In 1999 it wasn't feasible to run a whole-of-web crawler in Australia, because of latencies crawling international sites. (The vast majority of web servers were overseas). Australian network traffic charges were also beyond our budget – CSIRO was charged $20 a gigabyte for incoming traffic. At that time we would have likely needed to download more than 10 terabytes a year – $200,000. In contrast, when we rented servers from RackSpace in the USA, they came with a free allowance of multiple terabytes a year per server!

10. We were late to the party at using anchor text — Google were already using it in 1998 and McBryan's World Wide Web Worm years before that. (But we (particularly Nick Craswell) did some nice research on it.) We were also slow to incorporate link analysis and other prior probabilities which were mostly irrelevant to TREC ad hoc.)

11. It took me too long to realise the advantages of document-at-a-time (DAAT) over term-at-a-time (TAAT) query processing in large scale search with short queries. In TREC, queries were long, the ad hoc corpora were small, and TREC people generally didn't care about speed, so TAAT was perfectly fine.

12. P@NOPTIC was far too slow to embrace multi-lingual processing (Unicode/UTF-8). Having built PADRE to work on the ASCII documents of TREC, I initially wasted time adding PADRE

[1] Our incremental crawls still visited every page, but tried to avoid downloading unchanged pages by examining headers.

[2] If I remember correctly, the *WebFountain* crawler developed by IBM Research (`https://dl.acm.org/doi/pdf/10.1145/371920.371960`) used 40 racks of machines!

support for a choice of single-byte character sets, before biting the bullet and converting PADRE to work entirely in UTF-8. Years and years after that we were still having to diagnose and fix character-set problems in other components of the system. Even when team members developed new system components, their code or the libraries it used often assumed ASCII. ☹☹

13. When Trystan Upstill joined Google he told me that they were required to design and engineer everything at Web scale from the very start. We had the tendency to build small-scale prototypes and then have to face the problem of re-engineering them to work at the required scale. The analytics dashboard was a classic case – I insisted that it should be able to handle at least a billion queries, but our first implementation fell far short of that.
14. By 2000 Google were saying that user interaction data was the strongest signal, but because of our intranet focus we never had enough to think about A/B testing or to use click data in learning to rank.

10.2 Why Didn't We Become an Enterprise Search Gorilla?

My dream throughout the P@NOPTIC/Funnelback years was to be able to improve the productivity and competitiveness of organisations by providing highly effective search of both their external facing web sites and their internal information resources.

It was very clear that a large number of organisations suffered from ineffective search. Web site search is typically so bad that many people leave the site to search on Google or Bing. That can lead to loss of business because a broader scoped search may lead the visitor to a competitor.

Behind the firewall, things are even worse. Numerous industry studies, notably several by Susan Feldman of IDC Corporation, found huge costs associated with ineffective enterprise search. Knowledge workers were frequently unable to locate the information they needed to do their jobs. Feldman also found a high incidence of information being expensively recreated because the original couldn't be found.

In the rest of this section, I try to outline some of the reasons why we weren't able to realise my dream. (I have previously written a conference paper (2004)[3] and a chapter[4] in the Baeza-Yates and Ribeiro-Neto *Modern Information Retrieval* textbook, which outline many of the technical challenges in Enterprise Search.)

Looking back, it's a matter of considerable disappointment that we so seldom had the opportunity to bring research to bear on the problem of delivering high quality search behind the firewall. There were so many practical challenges associated with getting behind-the-firewall search operating at all!

10.2.1 Fragmentation of Purpose

Because P@NOPTIC, then Funnelback, grew organically, our sales people were forced to chase revenue wherever it could be found. We tried to address a large number of different markets: run-of-the-mill external web sites, behind-the-firewall enterprise search, e-commerce, call centre automation, whole-of-government search, university course finders, expertise finding, search over voice recordings, e-discovery, compliance and accessibility auditing, intelligence gathering[5], and embedded search.

The downside of this opportunism was that we didn't achieve alignment of R&D, marketing and sales. Products were developed and never sold, we bid for off-track opportunities which took R&D effort away from the main roadmap, and we could never come up with marketing messages which captured the essence of what we did.

As noted earlier, I eventually proposed that we focus on specific market segments that we knew well, creating product capabilities to address the needs of the segment, crafting marketing messages

[3] `https://david-hawking.net/pubs/hawking_adc04keynote.pdf`

[4] `https://david-hawking.net/pubs/ModernIR2_Hawking_chapter.pdf`

[5] Ben Pottier says, "The UK team engaged with a government entity to supply secure search across multi-million document domains for several partner institutions. After a year in build, the project went successfully live with happy clients."

and delivering them to the right channels for the segment, doing mass business development and lead generation for the segment, and recruiting sales people with credibility in the sector. My first candidate segment was universities and, as we have seen, it was pursued with some success after my departure.

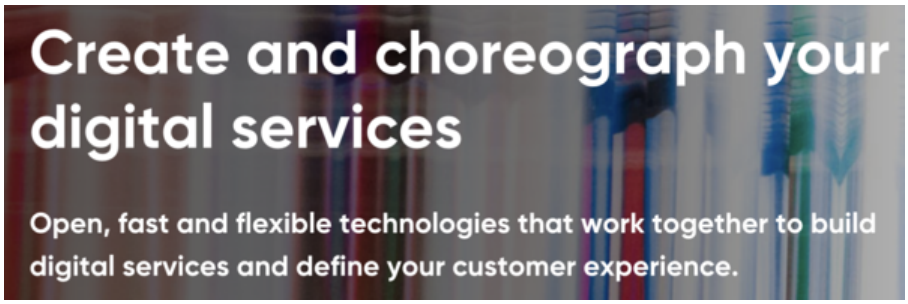
10.2.2 Clash of Business Models

My vision, originating in CSIRO, was to minimise the amount of human effort needed to be spent on each client. I thought we should strive to develop on-premise products and hosted solutions which were very easy to set up and configure, and easy to maintain and manage. I wanted to create bullet-proof products and services, and set the price to achieve large-scale market penetration.

In contrast, Squiz's business model was based on billable hours. Client sites would be designed and built from scratch with full project management. In Squiz UK, and when Squiz Australia started selling Funnelback, the clash of business models became evident.

10.2.3 Squiz Focus

Squiz's focus was and is on the presentation of information and the optimisation of stakeholder experience.



The main message on the Squiz homepage `https://www.squiz.net/`: 28 August 2021

Squiz were able to expand the reach of Funnelback technology on externally facing websites and applications, but they couldn't help us gain traction behind the firewall, where our major competitors over the years (Verity, Autonomy, FAST) made millions. They didn't have customer projects involving fileshares, repositories, data vaults, and records management systems.

10.2.4 Difficulties in Maintaining an Effective Sales Team

No matter how good a product is, and no matter how effective its marketing, company survival depends upon salespeople generating revenue. The large rewards paid to sales people sometimes irk the technical people who created the product being sold, but the product has no commercial value unless it's sold, and the volume of sales is dependent on rewards paid to the people who sell.

Having a highly effective sales team is a major determinant of company success but highly effective sales people are hard to find and recruit.

> **Q:** What's the difference between a computer salesperson and a used-car salesperson?
> **A:** The used-car salesperson knows when they're lying.

That joke from the 1980s reminds us of the poor regard in which salespeople are often held. But my limited experience trying to sell P@NOPTIC and Funnelback has given me an appreciation of their importance and of the difficulties they face.

Sales people talk about the sales funnel. At the top of the funnel you generate a large number of leads and gradually qualify them into prospects with greater probability of success. Finally, at the bottom of the funnel you try to close deals and put money in the bank. Bitter experience told us that you should never open the champagne until you've received the money.

We learned that certain salespeople were comfortable operating at the top of the funnel. There were some salespeople who actually loved being on the phone all day, cold calling prospects in the hope of generating leads. Some were very successful. There were others who generated leads by attending conferences and events, either presenting on search topics or by mingling with audiences, or staffing a booth promoting the product. You could call this business development – either promoting the specific product or promoting the value of effective search, regardless of vendor, in order to grow the market.

At the bottom of the funnel were the people who engaged in detailed negotiations with serious prospects, hopefully closing deals with good financial outcomes. It is a rare skill to be effective in this role.

A successful deal closer must fully understand the problem the customer wants to solve and the constraints under which they are operating. In a good sales meeting, the salesperson will likely spend more time listening than pitching.

A successful deal closer must understand when the prospect will be able to make a financial commitment and whether they are motivated to achieve the lowest cost or rather to obtain maximum value from a fixed amount of money. A very valuable piece of sales information is the prospect's budget for the project. The prospect may or may not be willing to reveal this.

A government agency may put up a case for funding the purchase of a search engine in the next financial year. If successful, there will probably be a fixed amount of money allocated. The agency will generally want to spend all of that amount on something which delivers greatest value and lowest risk. A proposal for only a fraction of the budget may be regarded with suspicion. Further, the person who argued the need for a million dollars when an effective solution can be obtained for a hundred thousand, may be thought a fool and have next year's budget reduced.

Closing a deal requires a full understanding of how the decision will be made within the organisation. Who will do the technical evaluation? Who will be involved in the negotiations? Who has ultimate sign-off? Who else will advise the final decision maker? Does Funnelback have a champion within the organisation? Is there a champion for a competitor? Does the organisation already have deep relationships with a major competitor? Early on, Funnelback used a formal *blue sheeting* process to make sure all these questions were addressed, and to avoid wasting time and money on prospects where success was highly unlikely. There were several examples when the process, formal or otherwise, wasn't followed, and huge optimism was eventually followed by devastating disappointment.

Another vital sales role is that of account management. It is a truism that it is far easier to retain an existing customer than to recruit a new one. A good account manager keeps in regular contact with customers, dealing with any problems arising during the year, and making sure that invoices are issued and paid on time. I understand that there was a period during which this wasn't happening in the UK.

Contact lists and schmoozing are sometimes important. Playing golf with high level decision makers or chatting with them in the Chairman's Lounge at the airport may ensure that your company is at least taken seriously when it comes to the final decision.

Measuring the effectiveness of business development people and lead generators is difficult, because you need to take into account not just the number of leads, but the size of the potential deals and the probability that they can be closed.

It would seem easy to calculate the effectiveness of a salesperson operating at the lower end of the funnel – Do they generate more revenue than the cost of employing them? However, a complication is that the time between generating a lead and closing a deal can take years. Brett Matson tells me of one university client who took eight years! Another complication is how to correctly value an ongoing deal. How do you compare a one-off $500k against $200k per annum when deciding on sales commissions?

Measurement of sales effectiveness is very important. You don't want to retain a salesperson who isn't earning their keep. Furthermore, sales people are inevitably employed on a retainer plus commissions and incentives. To be maximally successful, sales incentives and commissions should align the financial interests of the sales person with the long term interests of the company. Setting

the rate of commissions and the incentive scheme is difficult, and at risk of creating jealousies or perverse incentives. At other companies, incentive schemes have sometimes given rise to dishonest practices.

Across P@NOPTIC, Funnelback, and Squiz we've seen a high proportion of people in sales roles, who didn't succeed to the extent they would have liked: People whose personality wasn't suited to the role for which they were recruited; People who didn't understand the product or who were unable to understand the nature of the prospect's problem or the technical context; People who told prospects we had capabilities that we didn't; People who just didn't have the secret personality ingredient that might enable them to persuade a prospect to sign on the dotted line.

Great salespeople are quite rare, and difficult and expensive to hire. Why work for a small, relatively unknown company when they could be working on huge contracts for a large multi-national?

10.2.5 Challenges in Setting Up Behind-the-Firewall Search

To set up a demonstration search on an external web site or set of such web sites is usually very easy – crawl the site(s), build an index, mimic the look and feel, set up a side-by-side comparator, do a bit of quality control, and Robert's your father's brother! That's not the case with multi-repository search behind the firewall.

Apart from the cost and difficulty of acquiring connectors for the repositories, a considerable amount of time is needed on the part of both the customer and the supplier to get the connections established. It may take many days to suck out enough data to demonstrate the search. There may be a considerable amount of customisation of the search interface before the value of Funnelback search can be demonstrated. There may be considerable disruption to the customer's operations.

All-in-all, behind-the-firewall demos are an expensive exercise for both parties, to be undertaken without guarantee of success.

Another problem is that of confidentiality of information. Many clients with great need for effective search hold highly confidential information. Funnelback staff would need to obtain security clearances and sign confidentiality agreements prior to setting foot inside the organisation. Once inside the organisation, there are risks that an enterprise search engine may expose secret information to people who are not entitled to see it. Such leakage may potentially be caused by Funnelback bugs, bugs in connectors, or by system administrative failures within the organisation. Access Control Lists (ACLs) appear to be difficult to properly set up and maintain. Funnelback sometimes exposed existing errors in ACLs which allowed the wrong people access to restricted information.

The only possible path to getting Funnelback installed at many large organisations is via an outsourcing provider. In some cases, all IT purchasing must go through the outsourcer. In other cases, a third-party acquisition is possible but according to Steven Arnold[6] it's extremely unlikely that an American government agency would take this route because they wouldn't believe a small company would be able to master the complexity of the systems their IT outsourcer has set up, without the involvement of that outsourcer.

For an outsourcer the possibility of offering Funnelback to a large client may be immediately excluded because of competition with their own products or those of existing partners. If not immediately excluded, the possibility is seen entirely through the lens of increasing profit for the outsourcer.

Finally, we encountered very few cases where a company or agency issued a tender or RFP for supply of a behind-the-firewall, multi-repository enterprise search system. That was a pity, because issuing a tender would de-risk the situation for us by providing an indication that the organisation has recognised the value of a solution and has budgeted the necessary funds. I'm not sure why we didn't encounter such tenders. Perhaps a trail of expensive failed projects by other vendors had

[6] A prominent and very colourful American search industry commentator who claimed he bought me lunch at Parliament House in 2008 to interview me for `http://www.arnoldit.com/search-wizards-speak/funnelback.html`. (He didn't.) On a subsequent meeting he told me that a very senior AFP officer had taken him kangaroo shooting near Canberra with a machine gun. He also told a story of sitting next to a airline passenger who died in mid-flight. He told the crew to ensure that the corpse's seatbelt was securely fastened, their tray table stowed, and their seat in the upright position.

significantly reduced expectation of value and increased awareness of risk.

10.2.6 History of Expensive Failed Projects

Funnelback entered the enterprise search space at a stage when government agencies and large corporations were becoming aware of a large number of very expensive (many more than a million dollars) enterprise search projects which delivered very poor results and were regarded as failures. Some of our major competitors were acquired for sums in the billions, but their value was soon written down by an order of magnitude.

10.2.7 How Much Does Ineffective Search Really Cost?

Despite the previously mentioned industry reports on the "high cost of not finding information", the costs don't seem to be front-of-mind in many organisations. Perhaps the costs seem over-stated, often they are hidden, sometimes they may be accepted as inevitable due to pessimism about enterprise search projects. Productivity gains due to effective search are difficult to measure. How often do employees search? How much time do they save when the needed answer is at the top of ranking compared to the second page of results? What alternative do they have if they can't locate documents by searching?

I have to confess that although I quoted the findings from the industry studies and referenced the studies, I never subjected the methodology in the studies to careful scrutiny.

In cases, like help desks and call centres, where the benefits of effective knowledge base search are more readily measurable, the savings from effective search may be unpalatable – some staff lose their jobs.

10.2.8 Lack of Connectors

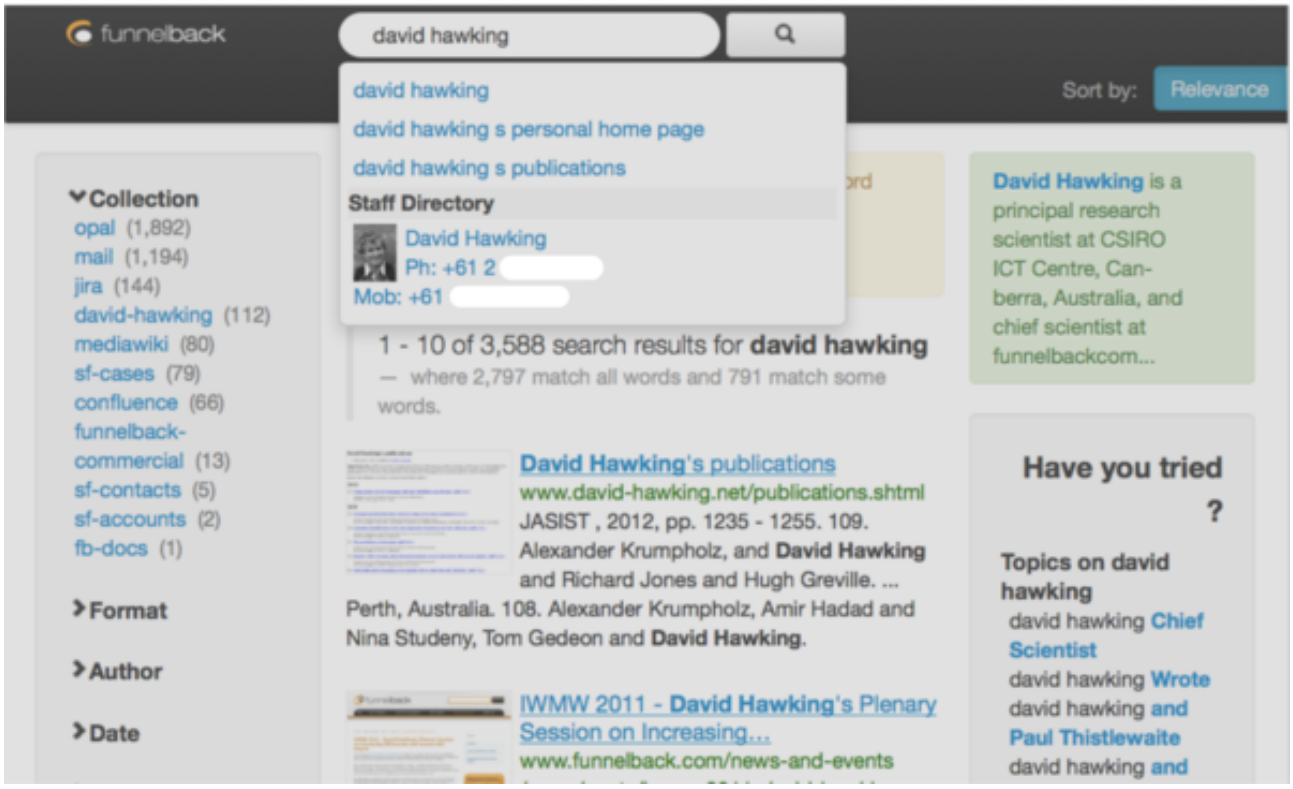
Behind the firewall, there is usually a plethora of different document repositories – e.g. Documentum, TRIM, OpenText, Objective, OneDrive, SAP, and dozens of others. An IT manager at Shell in Singapore told me that they had more than 50 repositories integrated into their FAST enterprise search system and each year departments competed to have their favourite repositories added to the service. At one stage Funnelback had as many different repositories (22) as it had members of staff. Challenges increase with the trend toward keeping internal information on external services such as SalesForce and Exchange 365.

To be able to search all the content, Funnelback has to connect to all the repositories, applying all the document level security access rules. Were Funnelback to attempt to develop its own connectors for every repository it was likely to encounter, it would need to maintain working installations of each of the repositories, and quite probably different versions of each. That would be intolerably expensive for a small company.

If Funnelback developed the connectors it would have to ensure that it accurately implemented the access controls. Any bugs in that code could result in a lawsuit.

Funnelback's larger competitors Verity, Autonomy and FAST, had the scale and the resources to develop their own connector libraries. In around 2000, Prabhakar Raghavan told me that it was the scale of Verity's connector library which gave it victory over its rivals. He said that bake-offs against competitors never got to the stage of measuring result quality – by the time the customer had got to the end of their list of required connectors, Verity was the last standing.

Funnelback partnered with a couple of different suppliers of connectors but there were always problems. It was always hard to agree on financial arrangements.



An early screenshot of Funnelback's internal enterprise search, showing: Facets (left) including match counts for eleven of 22 repositories; Segmented query auto-completion (top) with detailed result from the staff directory; Enterprise knowledge extraction (top right); Fluster (bottom right); and repository-diversified search results (centre) with document previews. It's a challenge to know how best to present results from so many repositories, some of which have popularity evidence and some of which do not.

10.2.9 Limits on the Possible

Enterprise search, particularly behind the firewall, is a very different problem to web search. In many search scenarios precision-laser-strike search is totally infeasible, due to the way information is organised and due to the lack of the signals on which web search most strongly depends.

Prior to the advent of successful web search, search was typified by an extended process of composing and refining a Boolean query, delivery of a large set of unranked results, and an extended process of scanning and vetting the results to find the required information. Some people, for example medical researchers and climate scientists, still apply this paradigm. To complete a valid meta-analysis or systematic review one has to be sure to have found all the relevant scientific papers. The queries submitted to PubMed or other literature databases must be recorded in the review or meta study.

The vast majority of searching these days is not like this. The rise of Google at the end of the previous millennium had a major effect on community expectation. People began to expect that their dramatically under-specified and often misspelled query would result in the perfect satisfaction of their information need within a fraction of a second. In a remarkably high proportion of cases of web search, that expectation is satisfied.

In enterprise search, particularly behind the firewall, these expectations are often impossible to achieve, because of the lack of popularity signals and the organisation of the information to be searched. Disappointment that behind-the-firewall search is never likely to operate "like Google" reduces corporate enthusiasm for embarking upon expensive and disruptive enterprise search projects.

10.2.10 Lack of Popularity Signals

The key to the Google revolution was the exploitation of popularity:

- Web author popularity in the form of links and anchor text; and
- Searcher popularity in the form of different types of clicks.

For a large proportion of web queries, it turns out that most people, or most people in a particular region, who submit a particular query expect a particular answer or set of answers. A negligible proportion of people submitting `Harry Potter` are trying to find the Harold Potter who lives in a beachside suburb of Melbourne.

There are user behaviour signals on individual web sites, both internal and external, but they are not as extensive or useful as on the web as a whole. Links and anchor text are usually there too but they are of less value because they do not provide a source of independent recommendation, and may have been produced semi-automatically by a content management system. In one case we found more than 90,000 links within an organisation, each with the anchor text of `Population` and each pointing to a page which was not the answer desired by 99.9% of people searching for `population`.

10.2.11 Depositing Rather Than Publishing

On the web many web publishers strive hard to have their content found. They frequently engage the services of search engine optimisers to improve their search engine rankings.

Behind the firewall, documents are typically *deposited* or *filed* rather than *published*. There is no effort made to improve searchability. Internal documents may be created in Microsoft Word on the person's home folder. They may be created in multiple versions, and some of them may be exported to PDF. PDFs (or Word versions) may be emailed to colleagues. Some of these versions may migrate to long term storage (e.g. a *data vault*). Eventually, at least one of the versions may be explicitly stored in a document management system.

When Funnelback tries to gather all the content, it may find dozens of copies and many versions. None of them are referenced by hyperlinks or anchor text. Any clicks resulting from previous searches may predominantly reference an older version or a form of the document already eliminated as a duplicate.

Brett Matson's thoughts on Section 10.2 above.

It's interesting that the failure to grow as big as FAST/Autonomy feels more for commercial reasons than technical (as compared to Section 10.1). Starting in Australia's tiny market with no viable springboard into other markets was a disadvantage. Squiz not allowing Funnelback to raise capital was a disadvantage. Not adopting your Higher Education strategy earlier was a mistake. We weren't as savvy as we could've been with all aspects of marketing (including the UI design). Perhaps there was also an element of caring too much about our customers, as opposed to some other companies' willingness to leave a trail of unhappy customers behind them.

10.2.12 Selling the Need for Enterprise Search

For many products like toilet paper, personal computers, and various categories of IT service, there is an established need, and companies have standing budgets for the purchase of the item. Vendors of toilet paper compete with each other on price, sustainability and quality – They don't have to explain what toilet paper is and what it can be used for, what benefits will accrue from supplying it to employees. At least one of the vendors is guaranteed to make a sale.

In contrast, Funnelback found that only a relatively small proportion of companies budgeted for acquiring and maintaining an enterprise search engine. To those that didn't, we first had to sell the likely return-on-investment from acquiring one. If successful, we then had to wait through one or more annual budget cycles while funding was secured, before being able to compete with other vendors.

10.2.13 Lack of Concern About Ranking Quality

Many people are vitriolic in their condemnation of poor search quality. As noted elsewhere, when search "fails", some write vitriolic letters to the Vice-Chancellor or the CEO. In the 1990s when many web search engines competed, people reacted with frustration at poor search quality, before switching en masse to an engine (Google) which demonstrably worked better than the others.

You'd think that this would cause the organisations purchasing an enterprise (or web site) search technology to try to maximise the satisfaction of employees or stakeholders using it, and to maximise returns to the business. Surprisingly, that's often not the case. Brett Matson has noted elsewhere that ranking quality was often not an important decision-making factor.

10.3 ✳Why Didn't Scientific Research Play a Bigger Role?

Funnelback would never have existed had it not been for the scientific research I and colleagues conducted at ANU, CSIRO and the ACSys Cooperative Research Centre. I believed that our research into understanding, measuring, and achieving effective search would lead to substantial competitive advantage over enterprise search companies not backed by such research. Large competitors like UltraSeek, Verity, Autonomy, Endeca and FAST seemed to fall into that category.

Scientific research at Verity

Prabhakar Raghavan was for a time Chief Scientist at Verity. He is an outstanding Computer Science researcher with a remarkably high h-index of 93. However, Verity didn't seem to take great advantage of his research expertise. As noted elsewhere, he told me that Verity bake-offs never came down to measurement of search quality. We found that Verity search quality was quite poor. Indeed UltraSeek, which Verity acquired from Inktomi in around 2003, seemed to provide better web site rankings than Verity's flagship product.

In the 22 years since the first P@NOPTIC production service, research into web ranking, query blending, query auto-completion, same site suppression, external annotations, and search quality measurement and tuning has played a critical role in delivering high quality search on web sites.

While I was an employee I was essentially the R in Funnelback R&D but a fair bit of what I did was routine engineering. Some of my work did involve adapting research innovations from whole-of-web search to the enterprise setting. To keep up with the state of the art, Funnelback very kindly allowed me to attend scientific conferences, supervise and examine PhD students, and participate in the Information Retrieval Facility in Vienna. Funnelback also allowed me to publish scientific articles, though I was careful to keep company secrets. We saw benefit in maintaining a reputation for innovation and scientific credibility.

Isn't search a solved problem?

In around 2000, Murray Cameron, Chief of CSIRO Mathematical and Information Sciences, brought a CSIRO Board member into my office and asked, "Why are we still doing research in Information Retrieval. Now we have Google, isn't search a solved problem?"

I find it much easier to argue forcefully when challenged, and responded, "there are many search problems and Google has delivered great answers to a few of them." I discussed this later with Trystan Upstill after he had been a key player in Google's Search Quality team for a year or two, and he basically agreed, saying that Google were still investing heavily in search research.

The Funnelback R&D team, typically comprising around 4 people, spent the vast majority of its time on routine software engineering, and very little on what would be considered scientific research. As the customer base grew, so did the necessity for bullet-proof software engineering and processes. Most new product features were software projects, usually emulating developments from elsewhere, and involving no theoretical innovations or solid backing in the scientific literature. Why weren't we able to invest more effort in real innovation?

One difficulty was the lack of focus. Initially, that was inevitable while we found the sweet-spot market segments, and while we generated the cashflow to keep afloat, but it meant we spent a lot of effort developing minimal viable product features for narrow segments which didn't really pay off. That may have been at the expense of really nailing the central things. That lack of focus was encouraged by the initial Funnelback board who expressed boredom with search, and encouraged us to look into new areas.

Brett Matson's CEO perspective

I was often asked by the R&D team why Funnelback wasn't investing more in research. I wished we could have, but I was never able to credibly justify it because, as an unfunded startup, employees were scarce and the list of urgent customer, sales, and product priorities was enormous (often in crisis due to customer deadlines and SLAs). Also, in the earlier days, Funnelback was well ahead of the pack in ranking quality, so ranking quality wasn't a blocker to growth. Later on, competitors slowly caught up (to varying extents), but ranking quality became less of an issue when it came to closing deals because customers often considered ranking quality a non issue.

For most of my time at Funnelback I was the only employee with a PhD.[7] In contrast, Bing was an organisation with more than 3000 employees, a high proportion of whom had PhDs. Many Bing innovations have been published in the scientific literature and/or patented.

Experience with government research funding

Funnelback was also a partner in a successful Australian Research Council Linkage grant involving ANU and the credit rating company Veda Advantage. The problem addressed was that of data matching or entity resolution in real time.[a] Is `Joe Bloggs of 32 Some Street, Elsware` the same person as `Joseph Andrew Bloggs of 32 Some St, Elsewhere`? I saw this sort of matching problem arising in future search applications and was interested to explore possible approaches. Unfortunately, the time overhead in the grant application and reporting processes was very cumbersome, and not justified by the benefit gained.

[a] `https://david-hawking.net/pubs/ramadan2013dmapps.pdf`

We were on the right track with our research. Many of the people who contributed to the research behind Funnelback were recruited directly into the engine rooms of major international web search companies. Trystan Upstill had a very successful career at the heart of Google search. Nick Craswell has had a similarly influential career at Microsoft's Bing (formerly Live Search) search engine, joined later by Peter Bailey, and then David Hawking and Paul Thomas. As Nick says:

> I think we demonstrated that a bunch of the expertise we developed was useful in whole-of-web search. Quite a few of us joined search engines (Trystan, Nick, Peter, Paul, Dave), and proved our good work on enterprise search (and research) by going on to make sustained technical contributions at web-scale engines. (Another example of that is Knut Magne Risvik.) These were all highly-specialized jobs, not the kind of job where you join the company and find out later what your technical area is. In my case, I'd been interested in IR ranking+metrics and was employed specifically to work on IR ranking+metrics.

Australia doesn't have a good record on private sector scientific research. Personally, it's a bit disappointing that the only scientific research (R) job created in Funnelback was my own.

10.4 But we did!

I felt that it was important in this chapter to critically analyse the path we took, in the hope that there may be lessons in our story for researchers, research organisations, and budding entrepreneurs who might be contemplating the commercialisation of IT research.

I now have time to think about what might have happened had different decisions been made.

[7] Dr Stuart Stephen and Dr Tim Jones were employed for relatively short periods.

- Should I have followed advice I received in 1999 to abandon CSIRO/ANU, create a new product in my garage and form my own company?
- Should I have pushed harder on the narrower vision of plug-and-play very-high-quality enterprise search?
- Should I have tried harder to control the destiny of the technology I founded?
- Should Fran, Nick, and I have mortgaged houses in 2000 and formed a startup?
- Should Brett, Stuart, and I pushed for a management buyout in 2009?
- Should we have tried harder to raise capital for R&D?

Counterfactuals alike these are very difficult to analyse, and the decisions were not mine to make alone. We became a brilliant team with a diversity of skills (and viewpoints). As you will see from the many comments in the next chapter, team members rate the team culture very highly.

And we were successful!

> I take enormous pride in the fact that our research led to the creation of a company which improved the quality of search at hundreds of organisations, earned many tens of millions of dollars in revenue, attained a peak of around 50 high-tech jobs, and significantly improved people's ability to find information.[8]

[8] Quoting from the introduction.

Chapter 11

"If I Could Say Just One Thing About Funnelback ..."

Brett Matson, long-time CEO: Everyone I speak to who spent time at Funnelback talks about it as one of the best places they've ever worked – strong camaraderie, innovation, excellent product, beating the big guys etc. It might not have been a billion dollar company, but it was definitely worth the effort.

Stuart Beil, long-time Funnelback chair: We can be immensely proud of what we achieved across so many financial, productivity, research, export, jobs and impact metrics. What I am most proud of is the teamwork, friendships, trust and culture we had developed as a company.... and for a brief moment in time we basked in the light of our own Camelot.

Francis Crimmins, one of the founders: Funnelback was an opportunity to work with smart people on challenging real-world problems.

Matt Taylor, FBUK GM: Overall I'd say my time at FB was amazing; it was fun, hard work, with the best colleagues and the most amazing tech to work with ... it made the whole experience not like a job but more like life!

Matt Sheppard, long time R&D Head: I really enjoyed working with the Funnelback folks over the years. The fact that all three Technical Directors at Instaclustr are former Funnelbackers must mean we were doing something right along the way! I was very proud of all the work the Funnelback R&D team put out in my time working there, but in the end I guess I feel like Funnelback never quite lived up to its potential overall, which feels a bit sad.

Shaw Xiao: The best thing about Funnelback was the people: Nicolas Guillaumin, Luke Butters, Dave Hawking, Matt Sheppard, Gordon Grace, Simon Oxwell, Brett Matson, Steve Barnes and so on. I am amazed that so many extraordinary staff were gathered there. They helped me and shaped me. It would not have been a great company without the people I have known there. After many years working with these people, I feel like they are a family. I was sad when I last walked into the Civic office, which coincidentally was the last day of moving the office. Everything was packed up, including all the servers that had been located there. Sadly most of the people I knew had left by then.

Prathima Chandra: Funnelback had a great working culture with everyone in the team willing to help out when needed. Even though we had remote teams we never felt out of the team or away as the teams would always look out for each other. Remote working was part of the culture way before COVID changed the world. I always feel proud of my journey with Funnelback and still fondly recall my experiences, learnings and just the passion of the team!

Will Parkinson: Funnelback was an excellent company to work for and I am very proud to have been a part of its history. I find it a little sad that all the people that built the company have moved on and the company itself seems to now be under resourced and a bit neglected now.

Nicolas Guillaumin, long-time member of R&D: Funnelback had a unique product, talented individuals and untapped potential. It was an incredible growth experience for me.

Mandhakini Iyer: Arriving in Australia as a first-generation immigrant, adapting to a new culture, finding a job, and settling in are challenging. But those I met at Funnelback (Brett Matson, Gordon Grace, Nico Guillaumin, Dave Hawking, Stuart Beil, Steve Barnes, Alwyn Davis, Matt Sheppard, Adrian Khoo, Natalie Grech, Narelle Bortolin, and many others) made me feel welcome and embraced me as part of family. Funnelback will always hold a special place in my heart.

Phillip Widdop, Funnelback UK sales: Whenever I think back to my time at Funnelback, I'm very grateful for what it has enabled me to achieve in my career. This was mainly because, at Funnelback, I was surrounded by much smarter people than me, who allowed me to grow professionally and, importantly, be myself.

Sonia Piton, Delivery Manager, Funnelback UK: I have never worked with a more brilliant, funny, and friendly bunch of people than the Funnelback family.

Jack Griffiths-Hawking: After joining Funnelback I was impressed by the awesome company culture, which Brett Matson and Narelle Bortolin were instrumental in fostering.

Luke Butters, R&D engineer, 2013 – 2021: The company was really fun to work at and I think we had a bunch of good engineers. For some time Funnelback was really a place that let you practice software engineering with a good degree of focus on solving interesting problems and autonomy. Funnelback really did let me indulge my appetite to make fast code. I would definitely do it all again.

Natalie Grech: The humans at Funnelback taught me that ego has no place in a productive workplace. When everyone wants everyone else around them to succeed, regardless of which arm of the business you're in, you receive support, respect and encouragement.

Appendices

Timeline

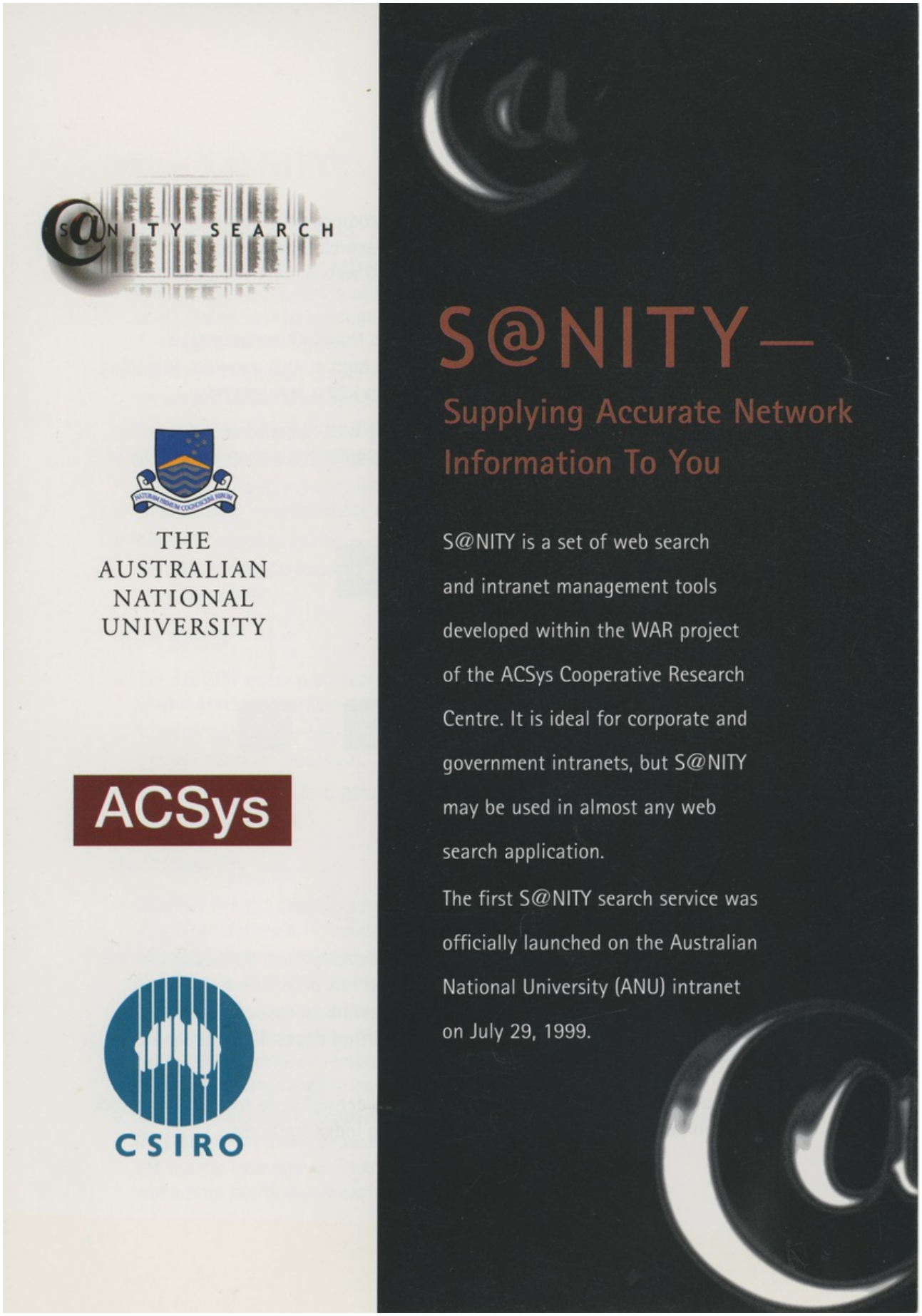
1991	David Hawking develops PADDY, parallel string search.
1994	David Hawking and Paul Thistlewaite participate in TREC-3 conference, using PADRE (developed from PADDY on Fujitsu AP1000).
1996	Nick Craswell commences PhD under supervision of Paul Thistlewaite.
1998	Paul Thistlewaite's WASPS and AWESOME proposals. Peter Bailey joins ANU to work on components of S@NITY.
1999	Paul Thistlewaite dies. S@NITY launched at ANU. Francis Crimmins joins CSIRO P@NOPTIC team.
2000	Nick Craswell completes PhD, later joins CSIRO. ResearchFinder is first paying P@NOPTIC customer.
2001	Synop becomes a P@NOPTIC reseller.
2003	Stuart Beil first becomes involved with P@NOPTIC.
2004	Nick Craswell leaves CSIRO for Microsoft Research Cambridge.
2005	Matt Sheppard joins CSIRO from Synop. AGIMO whole of government search contract commences. Funnelback Pty Ltd spins off, with Stuart Beil, Brett Matson, and Francis Crimmins as founders. Offices rented at Epicorp on Black Mountain. Matt, Peter Thew, and David Hawking (50%) seconded to Funnelback.
2006	Matt Sheppard joins Funnelback.
2007	Brett Matson becomes Funnelback CEO.
2008	David Hawking joins Funnelback. LSE becomes a Funnelback Client..
2009	Funnelback moves to Challis St Dickson. Funnelback acquired by Squiz.
2010	Funnelback UK joint venture set up..
2012	Funnelback and Squiz Canberra move into shared office in Allara Street.
2013	David Hawking and Stuart Beil leave Funnelback.
2016	Funnelback USA commences operation.
2017	Funnelback named *Visionary* in the Gartner Magic Quadrant.
2018	Funnelback signs Southwest Airlines, first high profile US customer.
2020	Brett Matson, Narelle Bortolin, Natalie Grech and others leave Funnelback. Funnelback is absorbed into Squiz.
2021	Matt Sheppard and Luke Butters leave Squiz.

Longest-serving Workers on P@NOPTIC/Funnelback

Counting service from the launch of S@NITY at ANU in July 1999 to October 2021. I.e. starting from the point where we had a commercialisable product.

Brett Matson	18 years	Feb 2002 – Nov 2020
Ben Pottier	14 years	Jun 2007 –
David Hawking	14 years	Jul 1999 – Nov 2013
Francis Crimmins	14 years	late 1999 – Jul 2014
Matt Sheppard	13 years	Jan 2006 – Mar 2008Jul 2009 – Feb 2021
Pete Levan	12 years	Oct 2009 –
Narelle Bortolin	10 years	Feb 2010 – Jul 2020
Luke Butters	10 years	Jun 2011 – Sep 2021
Stuart Beil	10 years	2003 – Dec 2013
Sonia Piton	9 years	Jan 2012 –
Natalie Grech	9 years	Jul 2011 – Sep 2020
Shaw Xiao	8 years	Jul 2011 – Apr 2020
Nicolas Guillaumin	8 years	Apr 2010 – Dec 2018
Phil Widdop	8 years	Sep 2009 – Aug 2017
Prathima Chandra	7 years	Feb 2011 – Aug 2018
Ben Tilley	7 years	Sep 2011 – Mar 2019
Gordon Grace	7 years	Nov 2010 – Dec 2017
Annie Pritchard	6 years	Jul 2007 – Mar 2014

S@NITY Brochure – July 1999

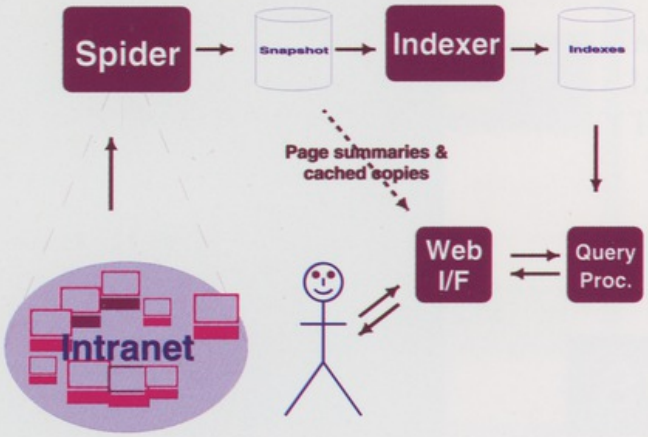


What is S@NITY?

The S@NITY search application consists of the components shown in the diagram below. At the time of launch, the ANU search service indexed approximately 170,000 pages served by over 130 web servers within the ANU domain.

Spidering took about 12 hours and indexing less than half an hour on a Pentium II system. When searching via the web, high quality answers, including query-biased summaries are typically returned in just a few seconds.

Webmasters within the ANU domain can provide host-specific or faculty-specific searches by forwarding user queries to S@NITY together with a list of servers to which the search should be confined.



Even Greater S@NITY

A number of S@NITY enhancements are under development, particularly in the area of intranet management tools, such as dead link detection and advice, monitoring of standards compliance, and general link analysis. For any page S@NITY will also be able to provide lists of referring pages. These tools will build on the spider snapshot and indexes.

ACSys is also interested in the use of mobile "spidering" code to reduce network and system load while maintaining or enhancing index freshness.

Why S@NITY?

HIGH QUALITY ANSWERS

- S@NITY uses the most effective relevance scoring formula currently known (Okapi BM25). The ability of this formula to deliver high quality result lists has been proven in international competitive evaluation (TREC).
  See **http://trec.nist.gov/**
- S@NITY is designed for use in a mixed metadata/content environment. Support for simple Dublin Core and Netscape-style metadata is built in.
- S@NITY provides high quality summaries of results. Sometimes there is no need to go to the original page.
- If a S@NITY query results in few exact matches (or none), results obtained by progressively weakening the constraints are presented as well.
- S@NITY Searches can be restricted to subsets of servers within the intranet to provide search services for individual web servers or for sub-groupings within the organisation.

HIGH SPEED

- The S@NITY query processor is fast. Typical web queries are processed over the ANU indexes in approximately half a second.
- S@NITY indexing is very fast (up to 13 gigabytes (2 million pages) per hour on a single Pentium processor).
- S@NITY features a fast, parallel, net-friendly spider capable of avoiding various forms of "spider trap".

HIGH CAPACITY

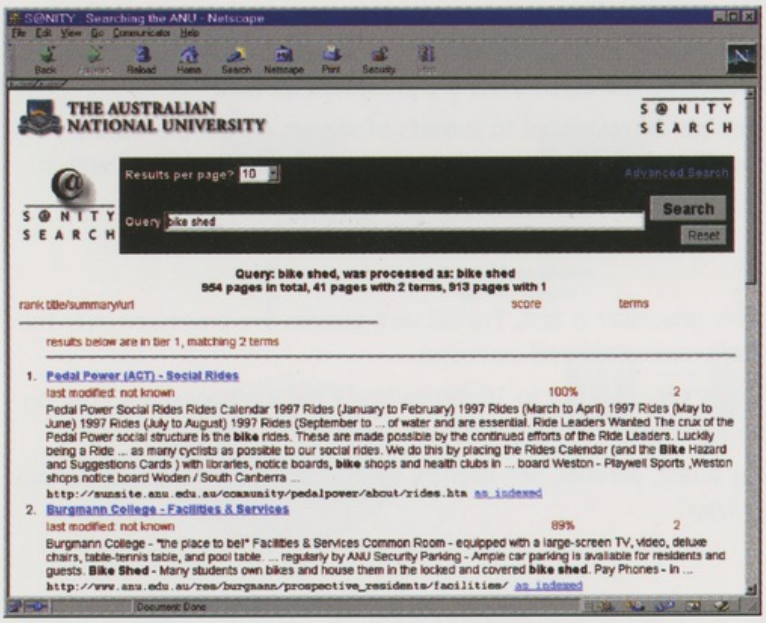
- S@NITY has the capability to handle very large intranets. A test collection of 18.5 million web pages can be indexed in 7.5 hours on a single processor.
- S@NITY indexing and query processing can take advantage of multiple networked machines to improve speed or data handling capabilities.
- S@NITY indexes are compact (Approximately 25% of raw text size in the normal case and as little as 5% when term positions are excluded.)

How can I find S@NITY?

The ANU S@NITY service is accessible at **http://search.anu.edu.au/**
or via the University's front page at **http://www.anu.edu.au/**

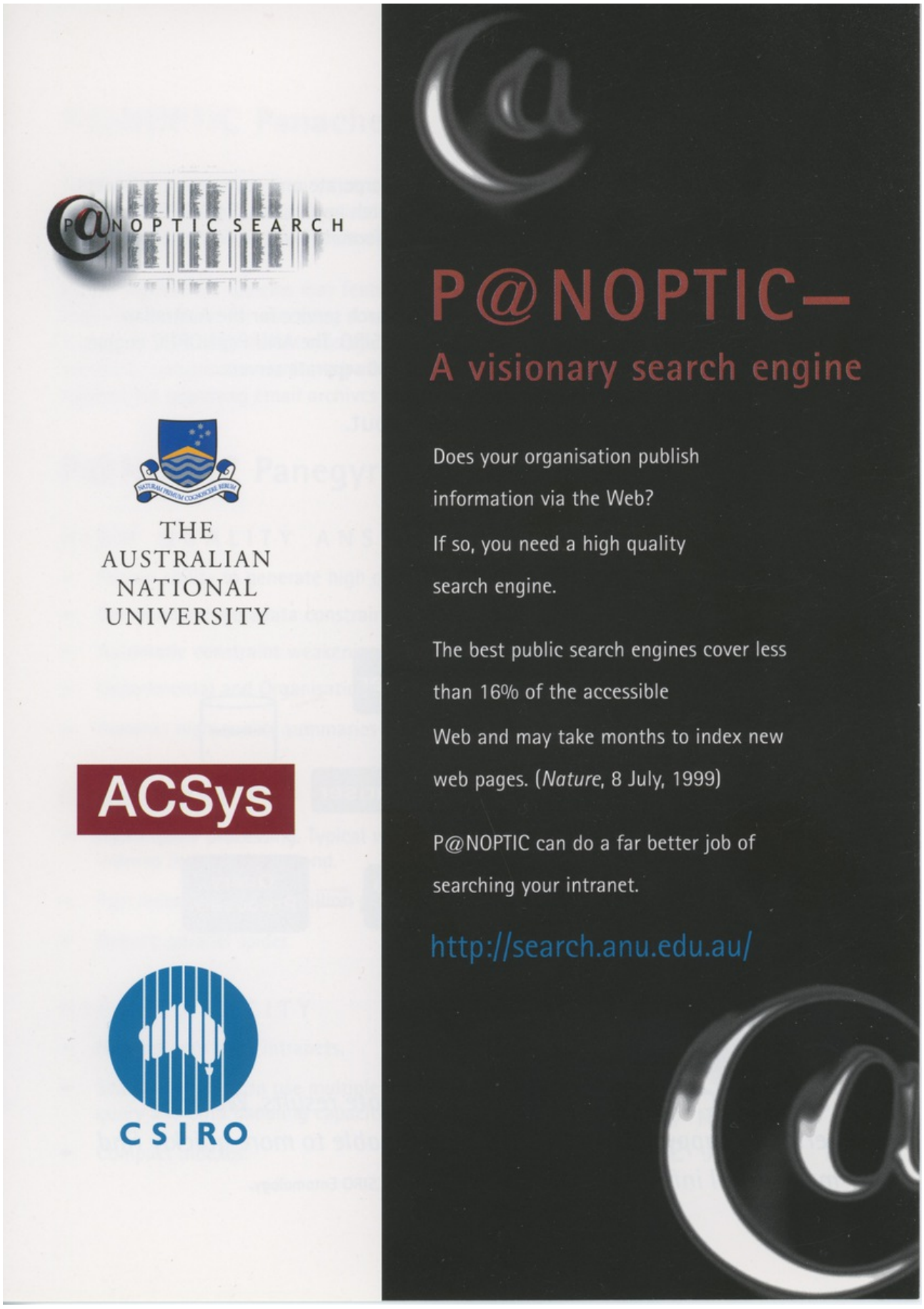
Further technical information on S@NITY is accessible via the search page.

CSIRO is currently considering adoption of S@NITY and discussions are under way with a number of other groups.



For further information about S@NITY, please contact
David Hawking
email: David.Hawking@cmis.csiro.au
phone: (02) 6216 7060
fax: (02) 6216 7111
web: http://pastime.anu.edu.au/WAR/

P@NOPTIC Brochure – Late 1999?



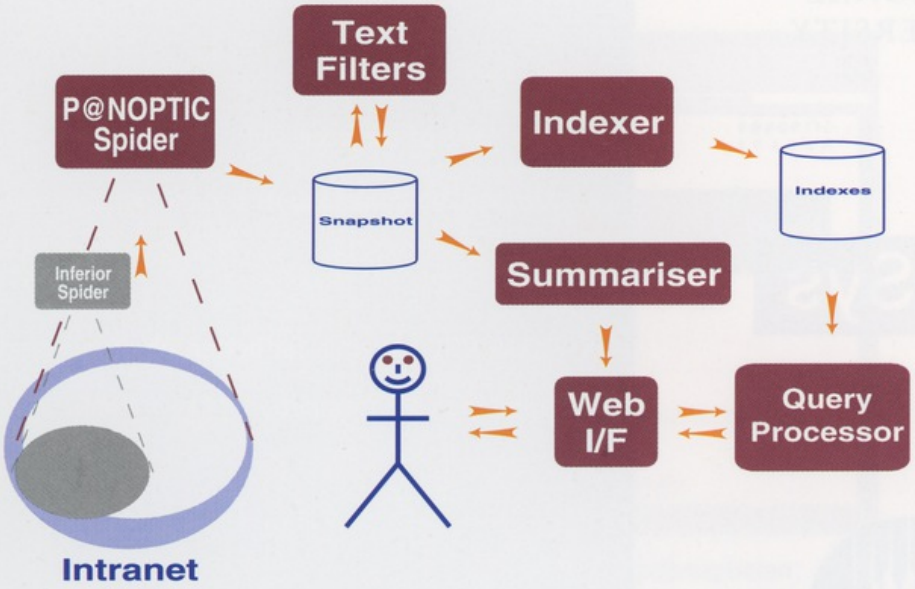
P@NOPTIC Panorama

The P@NOPTIC search engine is oriented towards corporate and government intranets and portals but may be used in almost any Web search application. P@NOPTIC supports Dublin Core and other common metadata formats and can be customised to meet the needs of your organisation.

A P@NOPTIC search engine provides the official search service for the Australian National University (ANU) and is also used within CSIRO. The ANU P@NOPTIC engine indexes over 300,000 ANU pages held by around 180 separate servers.

Visit http://search.anu.edu.au/ to try it out.

P@NOPTIC Pantechnicon



"P@NOPTIC has provided fast and accurate results. We are certainly happy with it. Visitors are now able to more quickly find meaningful information." B.Arrold, IT Manager, CSIRO Entomology.

P@NOPTIC Panache

The P@NOPTIC team has an international reputation in Information Retrieval areas including: Web search evaluation; test collection development; efficient and effective retrieval; and distributed IR (metasearch).

Future P@NOPTIC systems may feature focussed crawling, metasearching, and improved methods for accessing documents to be indexed. Also under consideration is a set of tools for intranet management, including dead link detection and advice, standards compliance monitoring and general link analysis. Prototype P@NOPTIC systems for searching email archives and local filesystems have been demonstrated.

P@NOPTIC Panegyric

HIGH QUALITY ANSWERS

- Proven ability to generate high quality results lists.
- Can combine metadata constraints and free-text query elements.
- Automatic constraint weakening reduces need to try multiple queries.
- Departmental and Organisational searches from the same index.
- Presents high quality summaries with query word highlighting.

HIGH SPEED

- Rapid query processing. Typical web queries are processed over the ANU indexes in around a second.
- Fast indexing (up to 2 million pages per hour on a single PC).
- Robust, parallel spider.

HIGH CAPACITY

- Handles very large intranets.
- Scalable design can use multiple machines to increase query and data handling capacity.
- Compact indexes.

P@NOPTIC Panoply of Products

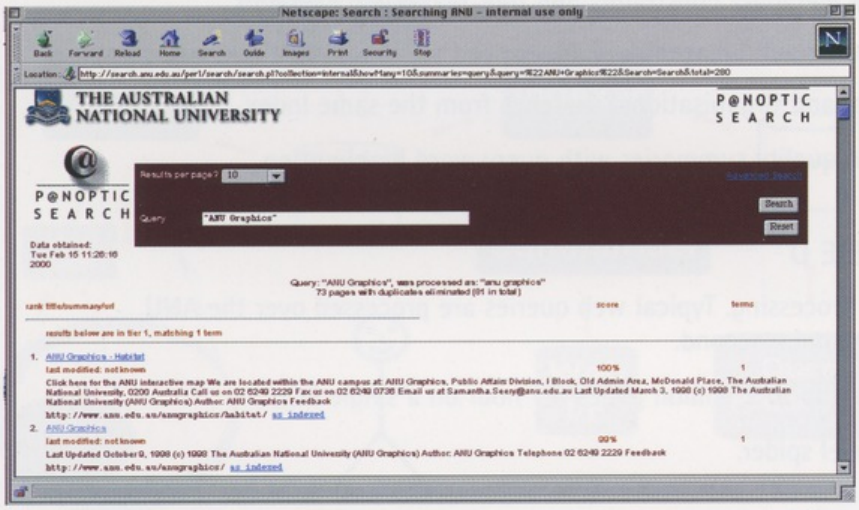
It has not yet been decided how P@NOPTIC will be commercialised.
The following "product" descriptions are therefore indicative only.

P@NOPTIC INTRANET SEARCH ENGINE

The P@NOPTIC Search Engine is a dedicated, low-cost PC designed to be plugged in to an organisation's intranet. Software installation and initial configuration is straightforward and subsequent administration is performed via a web interface. P@NOPTIC will soon includes open source filters for extracting text from common non-HTML formats such as Word and PDF. Customers would purchase the search engine and pay an annual fee to cover software licence and maintenance.

P@NOPTIC BUREAU SERVICE

A dedicated P@NOPTIC engine would be overkill for operators of small websites.
A bureau service may be offered in which a central P@NOPTIC search engine provides search services for multiple organisations. Customers would pay a monthly fee based on volume of data, frequency of spidering and query load.



If you would like to see more, please contact your p@noptician:

David Hawking
email: David.Hawking@cmis.csiro.au
phone: (02) 6216 7060
fax: (02) 6216 7111
http: //pastime.anu.edu.au/

http://search.anu.edu.au/

Proposed P@NOPTIC Pricing Schedule For Approval by ANU

David Hawking 6 Dec 2000

On-Site Search Engine Pricing

Prices subject to 10% GST.

Table 1: On-site Government/Corporate Pricing (per annum, pre-GST). Machine lease is optional and includes on-site maintenance in areas where the hardware supplier offers this service. The supplied machine is sized to support the specified number of documents with a maximum query processing load of around 4 million queries per year (ie 12 thousand per day). Otherwise customer may supply hardware to P@NOPTIC specification and self-install. In the event that CSIRO/ANU cease to support P@NOPTIC, either directly or through third parties, customers may negotiate with CSIRO/ANU to continue to use the product indefinitely in self-supported mode.

No. documents	P@NOPTIC lic.	S/W updates	Support	Total	Plus h/w lease
up to 100 000	$5 000	$ 5 000	$3 000	$13 000	$1 500 per m/c
up to 250 000	$7 000	$ 5 000	$4 000	$16 000	$1 500 per m/c
up to 500 000	$9 000	$ 5 000	$5 000	$19 000	$2 000 per m/c
up to 1 000 000	$11 000	$ 5 000	$6 000	$22 000	$3 000 per m/c
up to 2 500 000	$13 000	$ 5 000	$8 000	$26 000	$3 000 per m/c
up to 5 000 000	$15 000	$ 5 000	$10 000	$30 000	$4 000 per m/c
up to 10 000 000	$17 000	$ 5 000	$12 000	$34 000	$5 000 per m/c
above 10 000 000	-	-	-	POA	-

Installation: $1000 plus travel at cost.

Table 2: On-site Academic Pricing (per annum, pre-GST). See notes on Government/Corporate table above.

No. documents	P@NOPTIC lic.	S/W updates	Support	Total	Plus h/w lease
up to 100 000	$2 500	$ 2 500	$1 500	$6 500	$1 500 per m/c
up to 250 000	$3 500	$ 2 500	$2 000	$8 000	$1 500 per m/c
up to 500 000	$4 500	$ 2 500	$2 500	$9 500	$2 000 per m/c
up to 1 000 000	$5 500	$ 2 500	$3 000	$11 000	$3 000 per m/c
up to 2 500 000	$6 500	$ 2 500	$4 000	$13 000	$3 000 per m/c
up to 2 500 000	$7 500	$ 2 500	$5 000	$15 000	$4 000 per m/c
up to 10 000 000	$8 500	$ 2 500	$6 000	$17 000	$5 000 per m/c
above 10 000 000	-	-	-	POA	-

Installation: $500 plus travel at cost.

Bureau Service Pricing

Prices subject to 10% GST.

Remote search service provided by CSIRO is invoiced annually and is based on the following:

- The average number of queries processed per month (Q),
- The number of index updates in the month (U), and
- The maximum number of documents indexed (D).

$$\text{Annual fee} = 12 \times \frac{(Q + U * D)}{100}, \quad \text{Minimum of \$1200}$$

Table 3: Examples of Corporate/Government Pricing for P@NOPTIC Remote Search service. (pre-GST)

No. documents D	Updates/mo. U	Ave queries/mo. Q	Annual Fee
1000	1	1000	1200
1000	1	10000	1320
1000	4	10000	1680
1000	1	100000	12120
10000	1	1000	1320
10000	1	10000	2400
10000	4	10000	6000
10000	1	100000	13200
100000	1	1000	12120
100000	1	10000	13200
100000	4	10000	49200
100000	1	100000	24000

Initial setup: \$1100.

Table 4: Examples of Academic Pricing for P@NOPTIC Remote Search service. (pre-GST)

No. documents D	Updates/mo. U	Ave queries/mo. Q	Annual Fee
1000	1	1000	1200
1000	1	10000	1200
1000	4	10000	1200
1000	1	100000	6060
10000	1	1000	1200
10000	1	10000	1200
10000	4	10000	3000
10000	1	100000	6600
100000	1	1000	6060
100000	1	10000	6600
100000	4	10000	24600
100000	1	100000	12000

Initial setup: \$550.

Pricing Principles

1. P@NOPTIC should be priced to be much cheaper than products aimed at the rich corporate market, such as Verity, Fulcrum and OpenText because such products:

   (a) are well-established,

   (b) include features such as knowledge maps, thesauri and filters for a very wide range of document types, and

   (c) are probably living on borrowed time as far as high prices are concerned.

   I have no precise figures for these products but I believe that for a government department, they are probably in the ballpark of $50-$100k p.a.

2. P@NOPTIC should be priced cheaper than products such as Google Gold/Silver Site Search, at least until we have established ourselves and proven superiority. Google Gold costs around A$40k p.a. for a single site, is based on the crawl data used to support the main Google Web search engine and allows up to 4 million hits per year. Our target customers will receive far less than this number (eg. ANU - 100k). Note however that Google Gold can't supply internal search.

3. Although there are various free search services, such as `htdig`, there is likely to remain a paying market for a high performance, supported product with associated services.

4. Pricing should be related to the estimated, periodically auditable scale of the search operation, measured in:

   (a) number of pages indexed (both on-site and bureau service),

   (b) frequency of update (bureau service only), and

   (c) number of queries supported (bureau service only).

5. Academic/research discounts of 50% should apply to all P@NOPTIC products, excluding those licensed, leased or purchased from third parties, eg. third-party text filters, hardware etc.

6. We would expect to review price levels after the first six contracts are signed.

The Cnawen Project Brochure – 2008

www.csiro.au

Cnawen: corporate knowledge management services

Your company's corporate knowledge is valuable and you need it to be managed. Staff change roles and valuable information is lost. Documents emailed to and fro lose continuity and may not get to the right person. Legal discovery is a cost waiting to strike. Cnawen is built to help.

Cnawen is built around an archive of your corporate email, documents and more. Through analysis and refinement of documents placed in that archive, Cnawen services can support position handover, relationship management, legal discovery and more.

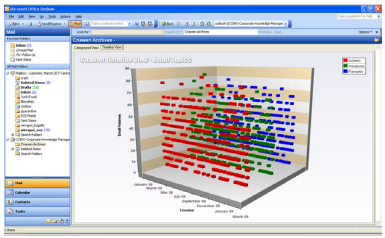
Relationship management

A vital task in any organisation is to manage and improve relationships. When talking to partners in other organisations it's important to know who else in your organisation has been talking to them and what the outcomes of those conversations have been. Furthermore, you need to know about their organisation itself.

Fast, easy briefings

Cnawen's relationship profiles start by giving you simple information about the organisations you are talking to. Useful, practical information such as their ABN, address and a summary of their business are all pulled together automatically.

Cnawen's profiles then go further to give you easy access to your company's previous dealings with the organisation. Regardless of age, if the document is in the Cnawen archive then it will be available in the profile. Profiles also show you who else in your organisation has communicated with the other organisation and what they said. Such information is invaluable in ensuring maximum benefit from future relationships.



> Cnawen's timeline view allows you to see who's talking about what

See the critical aspects

Conversations conducted over email can often be verbose and take some time to reach a conclusion. Such diversions can be particularly painful if you're reading from a mobile device. Cnawen can accurately search and summarise the dialogue and reveal the parts that you need to know.

Deliver on time

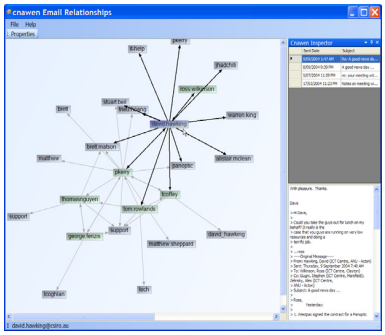
Commitments can be made easily with email, but can be hard to track. Cnawen can help keep both the commitments you make to others, and theirs to you, in order, referencing the original email conversations and documents.

Understand perspectives

Responding quickly to an email from an upset customer can make all the difference to your relationship. Using sentiment analysis, Cnawen can prioritise email based on how it is written. Less pressing email can be ranked lower, while important email gets through.

Handover support

Having staff start in a new position is hard enough without losing their predeceessors' contacts, documents and important agreements. Cnawen



> Cnawen's relationship viewer shows who's talking to whom

can let your new recruit find the long term contacts and decisions they need now, helping improve continuity in your organisation's relationships and commitments.

Easy records

Cnawen's easy archiving of records, such as email with attachments, means that even if the predecessors' mailboxes are gone, their project's email and attachments are still available.

Reveal the key messages

Using Cnawen's summariser and search services, staff can more easily see the important aspects of discussions that are likely to affect your new role.

Commit and deliver

Cnawen's ability to detect commitments in predecessors' email helps reduce action items slipping 'through the cracks' and resulting in disappointed customers.

Legal discovery

In recent years, organisations have faced large costs when courts have ordered them to reveal electronic records.[1] Most organisations regard themselves as underprepared.[2] Cnawen's managed knowledge makes the legal discovery process more reliable and easier.

Archive intelligently

The risk of litigation leads to many organisations simply keeping all documents, rather than only those that are likely to be required should legal action be taken. Cnawen gives your organisation the chance to archive documents that are likely to be required, greatly accelerating legal discovery. Well-archived, retrievable email may be also be critical evidence necessary to prove your organisation's prior art.

Useful day-to-day

Cnawen's archives are also useful in your day-to-day work.

Tagging

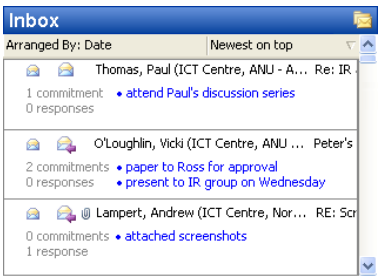
Tagging is used widely on the web, on such sites as flickr, to organise content flexibly and in a fashion where the user is in control. Cnawen allows you to tag your email in the same way. You can choose to share those tags with others, allowing a shared vocabulary to develop across your team and even organisation, further promoting well organised email.

Project drop box

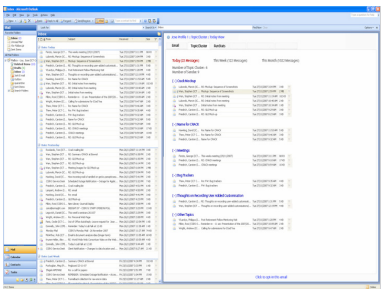
For any project involving more than one person, with priorities changing and people moving, there is a time when someone misses out on an email. With Cnawen's 'drop box' email archiving, everyone can find email, and their attachments, every time they need to. Drop boxes are searchable and archived, complete with attachments, which are easily backed up and will never be lost again.

Real world

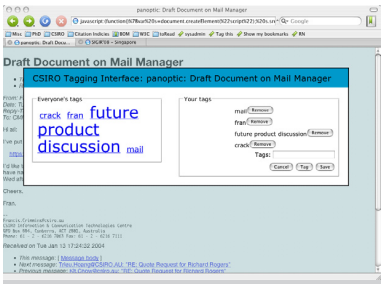
Cnawen is a service available for license from CSIRO.



> Cnawen's intelligent summaries can find commitments in email, showing what you really need to know



> Cnawen can cluster email so you can clearly see what you need



> Tag resources across your organisation for retrieval and sharing

1. White, Steve 2000, Discovery of Electronic Documents, *Digital Technology Law Journal*, vol. 2 no. 1
2. Fulbright & Jaworski 2006, *Annual Litigation Trends Survey Findings*

Contact Us

Phone: +61 2 6216 7019

Email: cnawen@cnawen.com

Web: www.cnawen.com

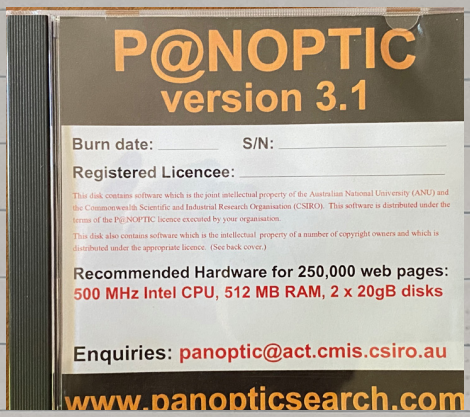
Your CSIRO

Australia is founding its future on science and innovation. Its national science agency, CSIRO, is a powerhouse of ideas, technologies and skills for building prosperity, growth, health and sustainability. It serves governments, industries, business and communities across the nation.

Index of People

Index of Customers, Prospects & Partners

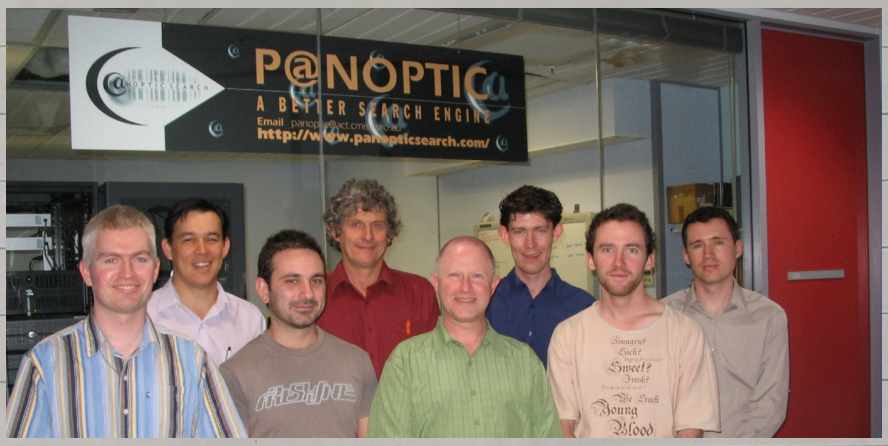
Index of Software & Services



When Google launched in around 1998 the world was amazed by its ability to locate even insignificant items among the profusion and chaos of the world wide web. Unfortunately, search engines deployed within organisations and on web sites at the time were often quite useless, leading to lost customers, frustrated stakeholders and unproductive employees. Keys to Google's success, including user interaction data, link information and anchortext, were sparse within organisations or more difficult to exploit. Furthermore, a search engine operating within an organisation rather than on the web is constrained to enforce that organisation's information access rules.

Australia's CSIRO saw an opportunity to address the problems of enterprise search by commercialising ongoing ANU/CSIRO research into text retrieval, creating the search engine known as P@NOPTIC which was spun off in 2005 as Funnelback Pty Ltd. The first site went live in July 1999 and the company was sold to Squiz in 2009.

Between 1999 and 2021, Funnelback grew to have representation in the UK, US, and Poland. It acquired hundreds of customers, including many blue ribbon banks and universities, and substantially raised the ability of their stakeholders to find things. Although growing organically, at its peak Funnelback employed more than 50 staff and has earned tens of millions of dollars of revenue. Not bad for a technology which started as an out-of-hours hobby.



Some readers will find value in this book as a study of successful commercialisation of research. Others may enjoy the story of the inspirational Funnelback team, the overcoming of technological challenges, and the roller coaster of thrills and disappointments inevitably associated with a tech start-up.

"Being part Funnelback meant being inspired by working with extraordinary people and being part of a team that had a growing realisation of what it could achieve, all while being under the shadow of, at any moment, a failure to compete bringing the swift end to everything we'd worked so hard on.

Against this exhilarating backdrop, Funnelback is the story of a diverse group of people who thrived on being the underdog and who fought to make an impact while living the uncertainties, complexities, daily challenges, and moments of happiness. Despite the improbable odds, these experiences ultimately shaped a team that brought a world-class product to a global market and successfully competed against the world's biggest companies."

Brett Matson, Funnelback CEO 2007 - 2022