Enterprise Search

What Works & What Doesn't

David Hawking

Nick Craswell, Francis Crimmins, Trystan Upstill

CSIRO Mathematical and Information Sciences

Canberra, Australia

Overview

- This talk makes a set of recommendations on how to set up high quality search of an organisation's key information assets. Some of them are nothing more than my personal opinion and may be ignored; many are thoroughly researched and should be taken seriously.

  You must decide which is which.

- I like to see organisational "intranets" as microcosms of the Web.

  If yours cannot be like that, this talk isn't for you.

References

- Hawking, Bailey & Craswell - CSIRO Tech Report
  - Efficient & flexible search using text and metadata
- Travis & Broder - Search Engines Meeting 2001
  - Web search quality vs. informational relevance
- Singhal & Kaszkiel - WWW10
  - A case study in web search using TREC algorithms
- Hawking, Craswell & Griffiths - WWW10 Poster
  - Which Search Engine is best at finding Online Services?
- Craswell, Hawking & Robertson - SIGIR 2001
  - Effective site finding using link anchor information

`www.ted.cmis.csiro.au/~nickc/pubs`

R1. Hide not your light under a bushel

Err on the side of caution ...

Publish everything!

If it’s not published, it can’t be found.

Seriously ...

- Don’t be excessively cautious with useful data
- Adopt a simple security model.
  - Open
  - Internal v. External, or
  - Secret v. Internal v. External

R2. Support "all" types of search

- What do your people search for?

1. Find key sites
   - library, HR department, psychology dept
2. Find named documents
   - IP policy, internal shuttle bus timetable
3. Find key information
   - scholarships, hiring, EEO policy
4. Find a person
   - CEO, Ev Brenner
5. Access services
   - Renew parking permit, Get a new staff card
6. Find relevant information
   - impact of trade practices act

R3. Use meaningful web path names for documents

1. The structure of your site(s) is useful to searchers/

- Consider the following ....

  `www.xyz.com/hr/policies/work_from_home.htm`

  `www.xyz.com/hr/policies/eeo.htm`

- Would you put these URLs on the noticeboard?
- Would you mail them to a colleague?
- Might you link to them?
- Which would you look in for the company's EEO policy?
- Where would you look for other HR policies?

- Good search engines exploit this type of info!

By contrast, ...

- XYZ inc "upgraded" their website. Now the URLs are:

  `www.xyz.com/x.y?page=000As0098-AAA-331210A0A111-7...`

  `www.xyz.com/x.y?page=000As0098-AAA-OA033121A111-7...`

- Would you put these URLs on the noticeboard?
- Would you mail them to a colleague?
- Might you link to them?
- Which would you look in for the company's EEO policy?
- Where would you look for other HR policies?

- There's nothing to exploit!
- (Like a library catalogued by accession number)

R4. Give your documents meaningful titles

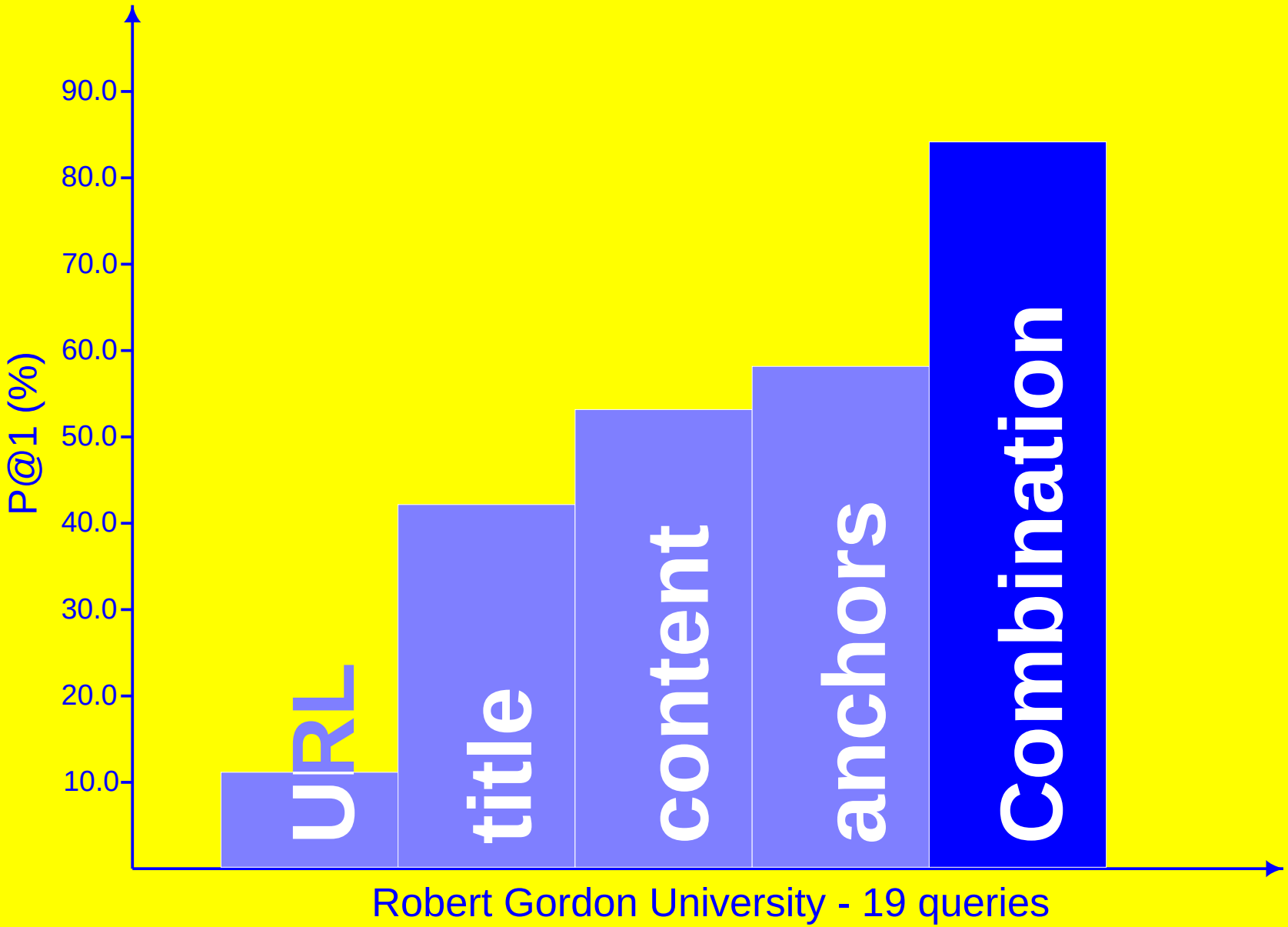
- Think about what someone might search for.
- Don’t re-use the same title on all the pages in a site!

R5. Link to your pages with descriptive anchortext

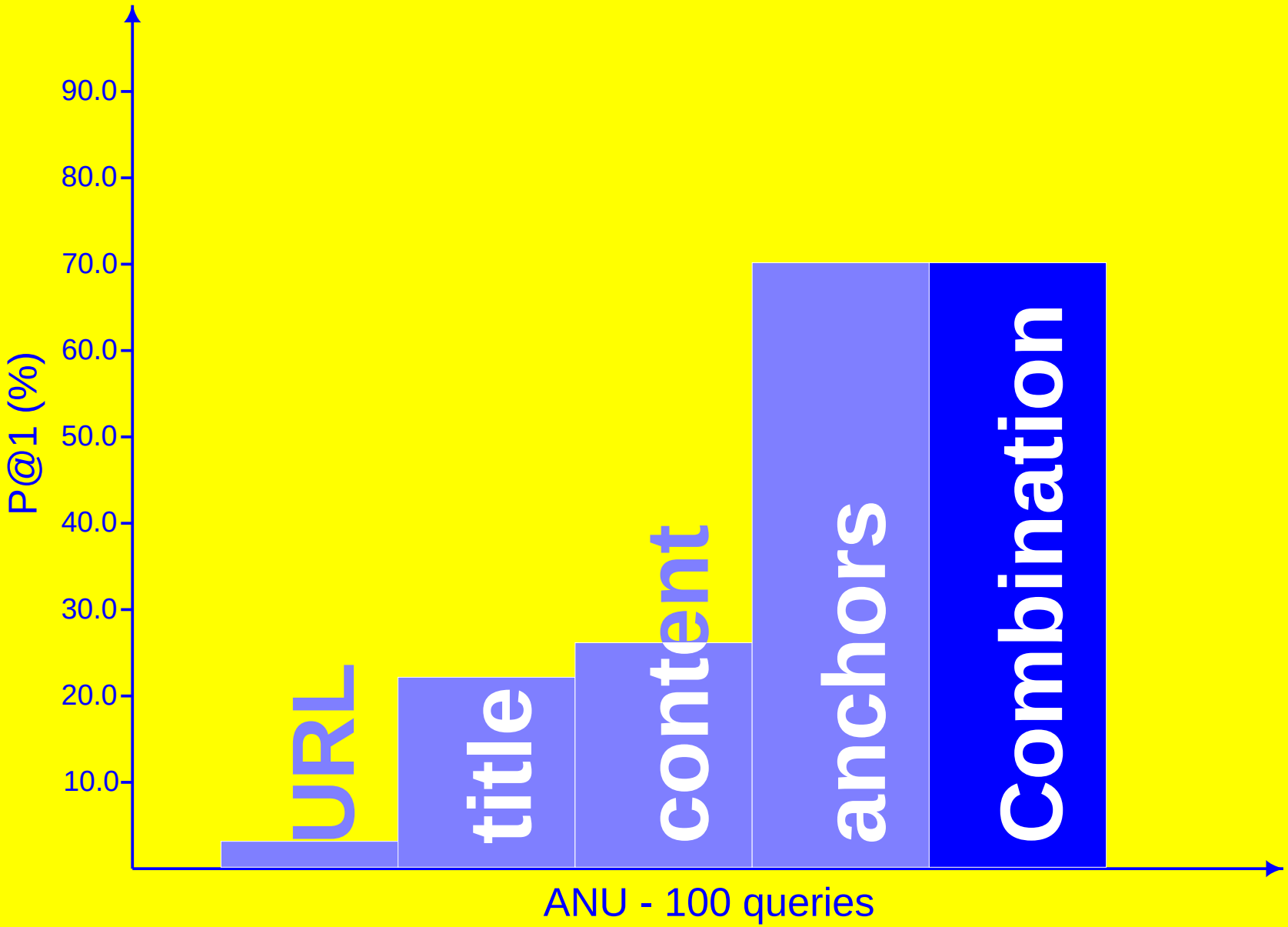
- Anchortext can provide rich descriptions of key pages.
  - Inlink count is not enough.
  - Enterprises are too small for PageRank
- Pages which are important because of what they are, not because of their content
- Frequency of reference doesn't confer authority.
  - Other pages may mention IP, copyright, intellectual property etc.
  - but that doesn't make them the company's official policy.

```
Human Resources homepage

Personnel

HR home

Personell

Click here
```

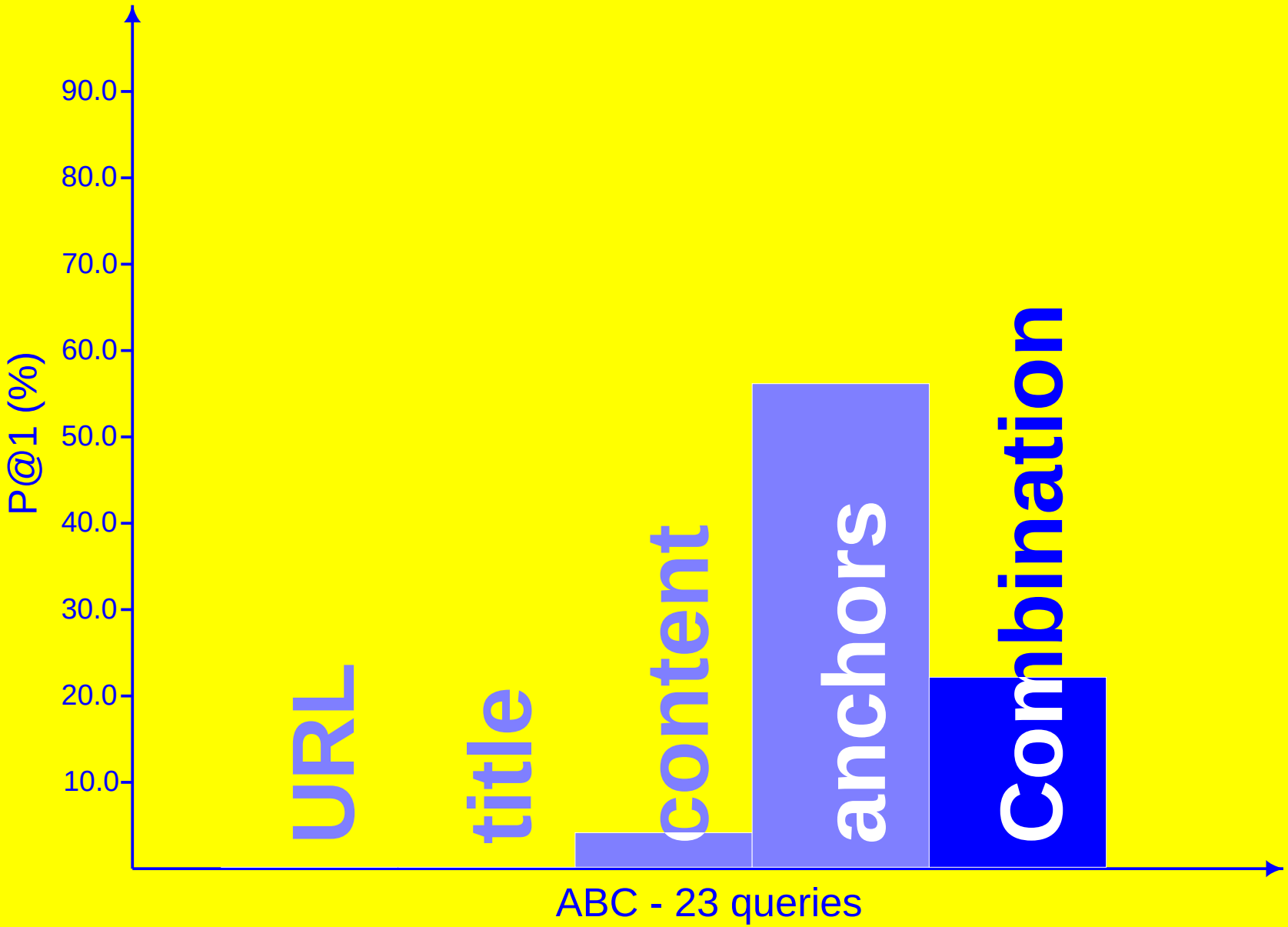
Results - Robert Gordon University



Results - ANU


P@1 (%)
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
URL
title
content
anchors
Combination
ANU - 100 queries

Results - ABC


P@1 (%)
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
URL
title
content
anchors
Combination
ABC - 23 queries

Is metadata useful in search?

- Obviously .... for information not easily derived from content
  - date, expiry date
  - author
  - publisher
  - audience
  - document type
    - ministerial directive
    - discussion paper
- Maybe ... for information about content
  - keywords
  - description
  - subject
- Useful for publishers, librarians, and archivists

Problems with metadata

- On the Internet
  - Can be used for spamming
  - Not enough of it
    - Important resources not discoverable
  - Inconsistent
- Within enterprises
  - Can cause resources to disappear
    - Missing metadata, or even better
    - misleading and inaccurate metadata
  - Authors copy metadata from one document to another!

How much metadata is there?

Enterprise	Author	Subject	Any DC
ACT Govt	12%	11%	49%
IP Aust.	41%	41%	45%
CSIRO intranet	22%	26%	26%
Other CSIRO	15%	18%	17%
ANU	11%	11%	4%
ABC	1%	62%	2%
USyd	4%	4%	1%
.GOV	6%	17%	1%
BluePages	4%	71%	1%
RGU	7%	10%	0%
UniNe	5%	14%	0%

R6. Find out what your searchers want

- Analyse common queries
- Do they give the right answer?
- If not,
  - How could metadata help?
    - Find efficient/effective ways of tagging
  - Query short cuts
    - Map queries/sub-queries to sites
    - Like paid query results
  - Would search quality be better in a narrower context?
- External metadata tagging by directory
  - How many "law" pages are there in

    `physics.anu.edu.au/`

    `www.anu.edu.au/forestry`

R7. Facilitate departmental search

- Publishers within your org. want to provide search of their stuff (only)
  - Search the HR department
  - Search servers in the department of computer science
  - Search the MS Office site
  - Search the mail archives
- 16,000 "departmental" search interfaces at the ANU
- Can implement via URL filtering, or
- By metadata constraints
  - Search only the President's edicts
  - Only need metadata on a small proportion of pages
- Potential loss of search quality

Multi-media documents - Outside resources

- You can search for these using
  - Metadata records
  - Anchor text
  - Image tags

R8. Don’t get carried away by intuition or tradition

- Preserve useful information when indexing
  - Index the metadata
  - Don’t stem words by default
  - Don’t throw away stopwords
  - Allow the possibility of case-sensitivity
  - Index multilingual data
  - disk space is plentiful
- Don’t do query expansion by default
  - Relevance feedback
  - Thesaurus expansion
  - Approximate matching
  - Stem matching
- Precision is usually more important than recall

R9. Get a good search engine

- Rate it on:

1. Data gathering
   - Crawling
   - Filesystem scanning
   - Database extraction
2. Text filtering
   - support lots of relevant formats?
3. Indexing / Query processing
   - Does it exploit all the evidence (see before)
   - Does it use a leading-edge relevance formula
   - Efficiency
4. Presentation
   - Result summaries?
   - Topic distillation?

R10. Integrate it with tasks & environment

- Wrappers to integrate with security model
- Can it integrate with day-to-day tasks
  - e.g. building approvals
  - e.g. immigration applications
  - e.g. journalism

RGU Demos

- Library
  - 1417 / 9343 docs contain the word
  - Title + anchors + url ---> rank 1
  - Anchors only ---> rank 2
  - Content-only ---> rank 27
- Hawking
  - Only one result ---> rank 1
- Hawking (with stemming)
  - Seven results ---> rank 3

Conclusions

R1. Hide not your light under a bushel

R2. Support all types of search

R3. Use meaningful web path names for documents

R4. Give your documents meaningful titles

R5. Link to your pages with descriptive anchortext

R6. Find out what your searchers want

R7. Facilitate departmental search

R8. Don’t get carried away by intuition or tradition

R9. Get a good search engine

R10. Integrate it with tasks & environment

Remember to break the Golden Rule

David.Hawking@csiro.au

www.panopticsearch.com