Introduction to WAIS the Wide Area Information Server

Peter Marshall, Manager: Academic Networking
The University of Western Ontario
Computing and Communications Services

<peter.marshall@uwo.ca>

Outline

- WAIS Concepts
- What's Out there?
- WAIS at UWO
- Searching with SWAIS
- Setting up Databases
- Problems
- Future
- For Further Information

Concepts: Overview

Concepts: Implementations

- Clients
  - Unix
    - shell commands
    - curses-based screens
    - X-windows
  - MS-DOS: packet-drivers (2)
  - Windows
  - Macintosh
  - VMS: Wollongong; Multinet; Digital's Ultrix Connection
  - "gopher"s of all kinds!
- Servers
  - Unix systems (all kinds)
  - Connection Machines
  - VMS
  - Macintosh (soon?)

Concepts: Who?

- over 335 servers
- more than 6,000 hosts
- greater than 10,000 direct users
- plus gopher gateway users
- world-wide
  - USA, Canada
  - UK, France, Germany, Switzerland
  - Australia, New Zealand
  - Israel
- (soon) major library catalogue software vendors

What's Out There?

- An index to journalism periodicals: **Does the reporting of John Crosby's speeches pose an ethical dilemma?**
- ERIC (Educational Resources Information Center) Digests: **Does taking notes help students in information retention?**
- Communications of the ACM (full-text): **Show me that recent article on user interfaces to computer programs.**
- The Gutenberg Project archives: **How about checking out a copy of Brontë's** *Wuthering Heights***?**
- Poetry archives: **Find me a poem about roses.**
- Human DNA sequences (updated as they are discovered).
- Weather satellite pictures (updated every hour): **Show me the current ground conditions in North America.**

- The KJV of the Bible: **Who was the guy with all the boils?**
- The Koran: **Is the Garden of Eden mentioned?**
- FTP'able README files: **Where would I find a program that converts scribe to TeX?**

WAIS at UWO

- Still experimental...
  - mostly unix: /ccs/export to sun3/4, mips
  - no VMS client at CCS
  - some trials with PC and Mac clients
- Local databases:
  - Index to Journalism Periodicals
  - Change Notices
  - Newsletters
  - Personal E-Mail

Index to Journalism Periodicals

- Bibliographic index: 40 journals on journalism
- Maintained by UWO Graduate School of Journalism
- 10 years cumulative data: >15,000 entries
- Fiche sold widely
- Experimental WAIS index: Widely used but...Will anyone pay?
- Used locally by GSOJ students
- other GSOJ databases possible: LFP, research papers
- changing the update frequency: 6 months $\longrightarrow$ 1 month

Change Notices

- help coordinate changes, spot problems
- local usenet news group
- poor retrieval mechanisms
- tried mh/xmh folders
- WAIS much better
- example
- On-line maintenance manual
  - Changing the way changes notices are written

Other UWO Databases

- Newsletters
  - CCS's *Focus*, InfoCentre's *Info Exchange*
  - Campus Computer Store Price list
  - *CA*Net Newsletter*
  - *LinkLetter*
- Personal E-Mail

Other Possible Indexes

- Senate and Board of Governors minutes.
- Campus newspapers: student and administration.
- Restaurant and movie reviews.
- Course outlines and calendars.
- Departmental minutes (using a private server).
- Usage write-ups Unix man pages and local manuals.
- Library catalogues.
- Frequently Asked Questions.
  - from usenet group news.answers
  - local FAQ one for/with the help desk

Searching with `swais`

English, no Venn, no booleans — easy?

1. Start up the querying program: `swais`
2. Choose database(s): cursors and ␣
3. Enter search terms: `w, ...,` ↩
4. View ranked "headlines"
5. Choose documents for retrieval: cursors
6. View or save document: ␣, `m` or `|`

Exercise # 1: getting started

1. Start up the `swais` program. (Type `swais`↩.)
2. Select the `journalism.periodicals` database for searching. (Move the cursor over it and press the space bar.)
3. Search using `Bill Clinton and Gennifer Flowers` and read a couple of the entries. (Type `w` and then the words above to the **Keywords:** prompt.)
4. Do another search using the keyword `Bucharest`. (Type `w` to get back to the **Keywords:** prompt, control-u to clear the line and then enter `Bucharest` ↩.) What do you get back?
5. Deselect the `journalism.periodicals` database. (Type `s` to get to the sources display. Move the cursor over the journalism entry and type a space.)
6. Exit from the SWAIS program. (Type `q`.)

Searching: Finding a Data Source

- menu on local server?
- directory-of-servers
  1. initial search of directory-of-servers
  2. choose appropriate database(s)
  3. add them to your "personal" list: `u` (use)
  4. use them in a search
- cache, but no invalidation
- database sources fairly stable

Exercise # 2: Add a Database

1. Make a directory to store pointers to databases in

   ```
   mkdir ~/wais-sources
   ```

2. Find the new (1991) CIA world fact book and add it to your personal list of searchable WAIS databases.

Searching: Multiple Databases

- local arbitration and sorting
- coarse grained parallelism

Exercise # 3: Searching Multiple Databases

1. Search in the `journalism.periodicals` and one of the World Fact Book databases (1990 or 1991) for information about **Singapore**. (Select both databases from the `s` (source select) screen, use `w` to get to the **Keywords:** prompt and enter `Singapore`.)
2. Mail an entry to yourself — if you have e-mail. (Move the cursor to one of the returned entries and type the `m` command. Then enter you e-mail address at the **Address:** prompt.)

A Personal Database: e-mail

1. Save the e-mail messages that you want to keep in mail folder files. (We'll assume that they are in your `./mail` directory.)
2. Path into ./mail and give the command

```
waisindex -nocat -mem 40 -r -d my-mail -t mail *
```

3. Copy the resulting `my-mail.src` file into your directory `./wais-sources`. Leave all the rest of the `my-mail.*` files around.
4. Use `my-mail.src` like any other WAIS database.
5. Periodically re-run the `waisindex` command to update the indexes as you add more mail messages.

Setting up a Public Server

1. Find the data
2. (buy more disk)
3. Massage the data or waisindex
4. (buy even more disk!)
5. Run `waisindex` on the data. Use append-mode for updates
6. Edit `.src` file
7. *Set up a Z3950 service (usu. port 210)
8. *Add a line to /etc/inetd to run `waisserver`
9. *Log file management procedures
10. Register by sending the `.src` file to **wais-directory-of-servers@quake.think.com**

Problems with WAIS

- Implementation Deficiencies
  - charging: charging your credit card
  - maximum number of documents returned
  - coding & recompile required to describe a new data format
  - searching model imprecise and often not adequate
    - No natural language
    - no word stemming or variants
    - no booleans
    - poor relevance feedback implementations
- Architectural Problems
  - directory scaling
  - X.500?

The Future

Bright!

Directions:

- More flexible and powerful searching:
  - stripping suffixes like "ing", "ed" and "s" for English databases
  - boolean searching
  - more powerful, cross-server relevance feedback
  - proximity search capabilities, phrasing
  - better understanding of natural language queries
  - non-word searching

The Future continued

- improved index density for space savings
- improved client implementations (especially MS-DOS)
- specialized, imbedded client applications
- Commercial databases, charging and access control. $$$!!!
- WAIS front-end $\longleftrightarrow$ PAT/Lector back-end
- Multi-media: sound, video, graphics
  - music retrieval
- compatible commercial Library catalog software

For Further Information

- See Peter Marshall's paper *WAIS: Wide Area Information Server or Anonymous What???*: via anonymous ftp on julian as `ftp/doc/wais-paper.ps.Z`
- WAIS bibliography: via anonymous ftp to **think.com** as `wais/bibliography.txt`
- Newsgroups: `comp.infosystems.wais` (general discussion), `uwo.comp.wais` (developers e-mail list)
- See Roberto Sierra's *WAIS & Music* in wais-discussion archives
- Check out gopher and WWW

Relevance Feedback:
Narrowing the Search

- Not really English queries
- not really ORs
- ranking
- using relevant documents to find others
- better than Booleans!