Web Retrieval:
Clues on IR Systems and Services Architecture

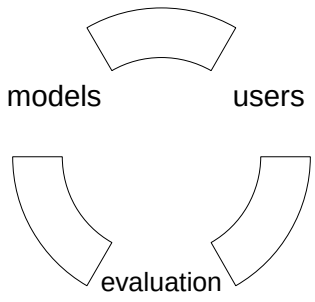
Maristella Agosti

Information Management System Research Group
Department of Information Engineering
University of Padova, Italy

European Summer School on Information Retrieval
3rd September 2003, Aussois, France

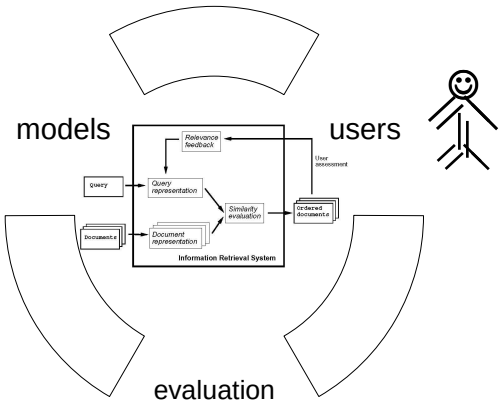
Contents of Former Lectures

- The scope of the initial days lectures has been to address all basic issues of modern Information Retrieval (IR) to set the scene of this complex and fascinating discipline:
  - models,
  - users, and
  - evaluation.



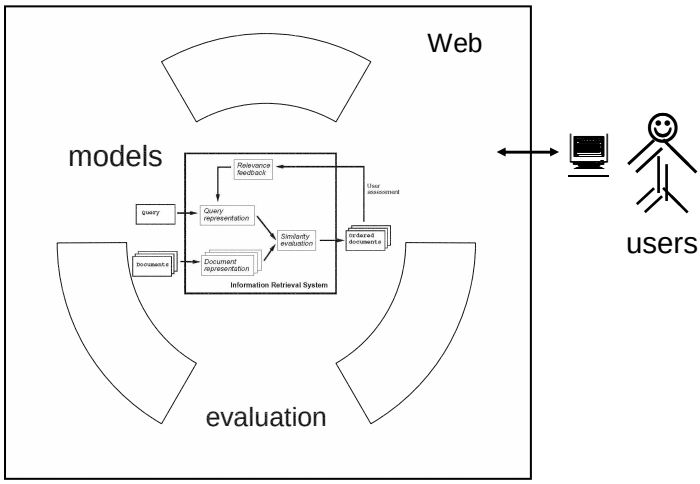
Scope of Present Lecture

- Main scope of this lecture is to give some clues to start to translate/transform these three aspects into effective and usable Information Retrieval Systems or Services (IRS).



IRS “on the Web”

- We consider only the case where the IR Systems/Services are “on the Web”, that is the IRS are accessible through the Web permitting Web Retrieval of information and documents.



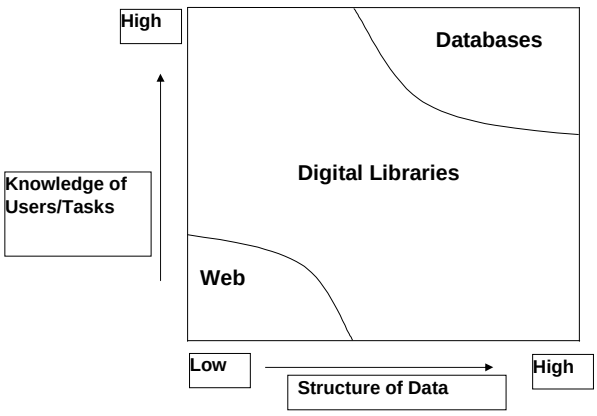
Focus

- Web Retrieval is a vast subject. Here we concentrate on the aspects that are related to the retrieval of:
  - collections of Web pages,
  - collections of digital documents, or
  - collections of catalogues data

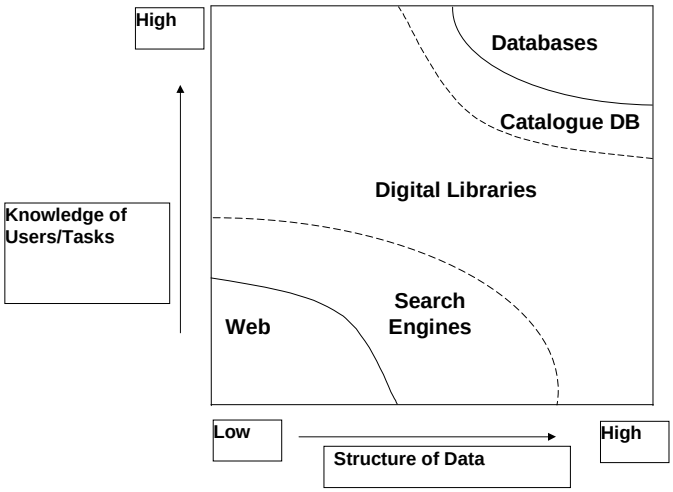
  for users accessing those collections through a Web browser.
- "IR System" and "IR Service": similarities and differences.
- Main IR Systems and Services of interest are:
  - search engines cooperating with search agents for the retrieval of Web pages,
  - search services usable through digital library architectures for the retrieval of digital documents, and
  - bibliographic/catalogue services for the retrieval of catalogue information and the localization of digital or traditional library material.

The Information Space Defined by DELOS

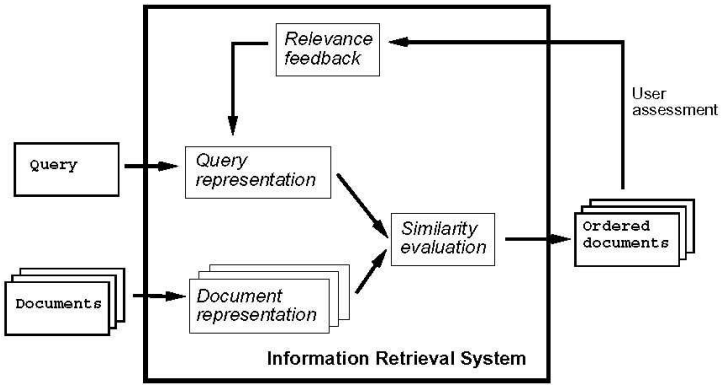
- From [DELOS, 2001]:



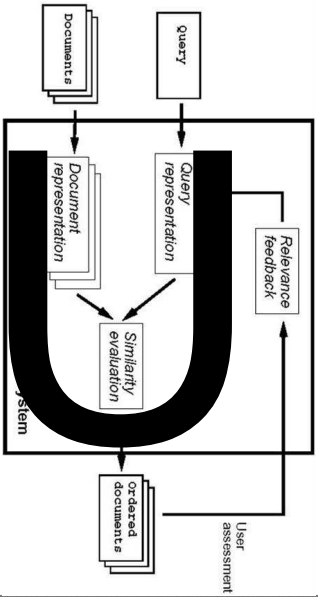
The DELOS Information Space Revisited



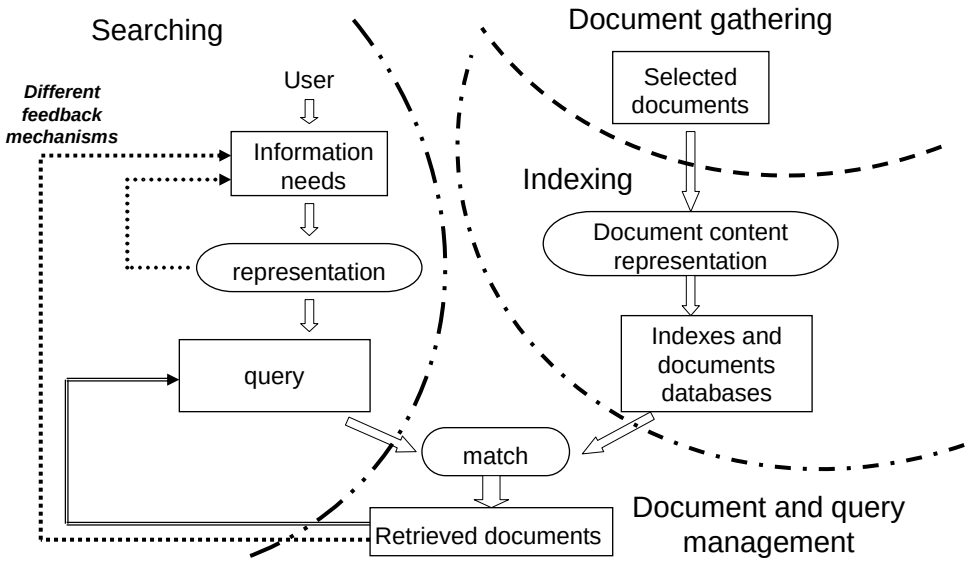
IR: Process and System/Service



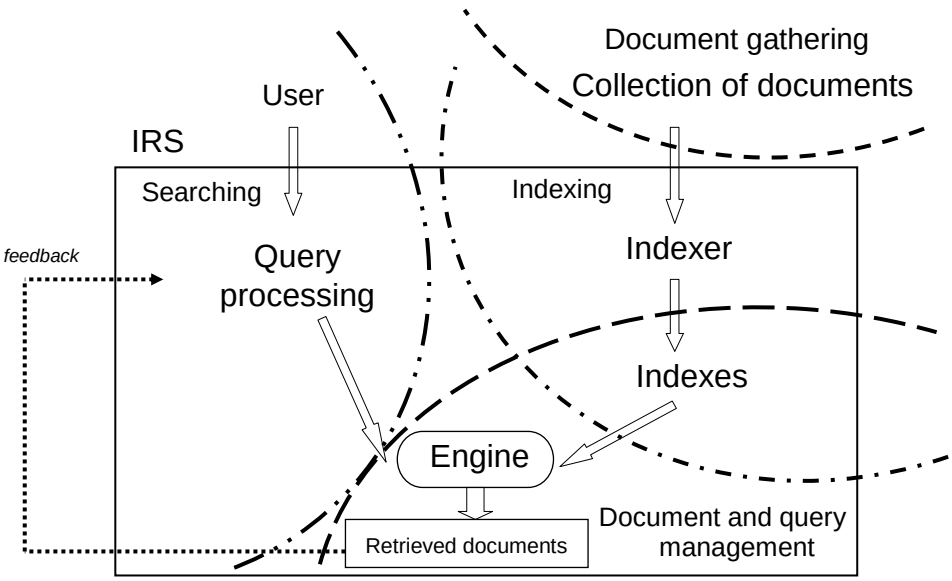
The IR "U" Process



Phases of the IR Process



Architecture of a Traditional IRS



Search Engines

Search engines cooperating with “search agents” and “file traverser” for the retrieval of Web pages and documents by direct search

(direct search versus navigation)

A Web Page as an IR Document

- A **Web page** corresponds to a **document** in traditional IR.
- Web pages are really different in size and in the type of files that can constitute them (e.g.: text, graphics, sound).
- Collection of Web documents:
  - In many cases the collection of documents that is processed by a search engine is a collection of selected Web pages/documents of interest in the part of the Web which is “directly” available;
  - this means that often are excluded Web pages and documents that cannot considered for direct indexing (e.g.: pages with authorization requirements, or pages excluded using the robot exclusion standard).

Document Gathering

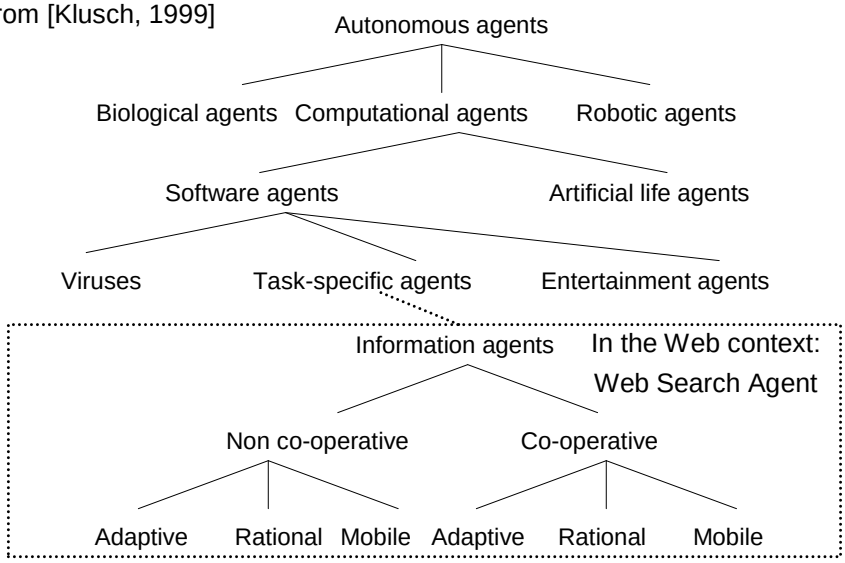
- Target:
  - construction of the collection of the Web pages and documents that form the universe of interest that the software tools have to index and manage.
- The user is going to search and retrieve “relevant” documents from this subset of the real Web.
- Output of this phase can be:
  - “virtual collection” (docs discarded after indexing),
  - “physical collection” (docs maintained), or
  - both.

SW Tools for Document Gathering

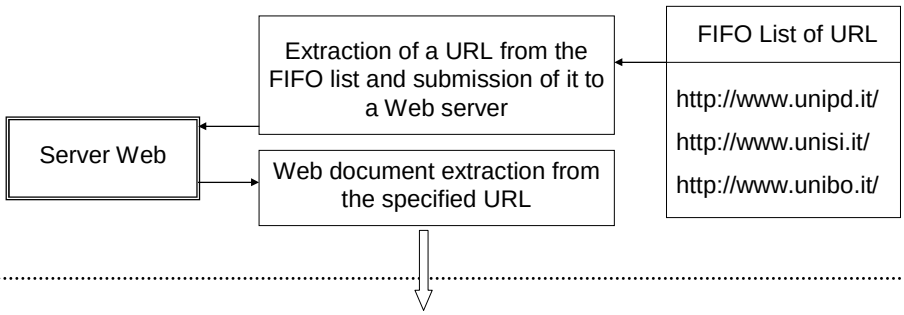
- The gathering of the documents for building the collection of Web pages can be done in two different ways:
  - the Web pages are submitted by users or companies to the search engine,
  - the Web pages are collected by a software tool, usually a Web Search Agent (WSA), where the WSA is an Information Agent which operates in the Web; see [Klusch, 1999].
    Typical used names for identifying a WSA are: spider, crawler, and robot.

A Classification of Information Agents

From [Klusch, 1999]



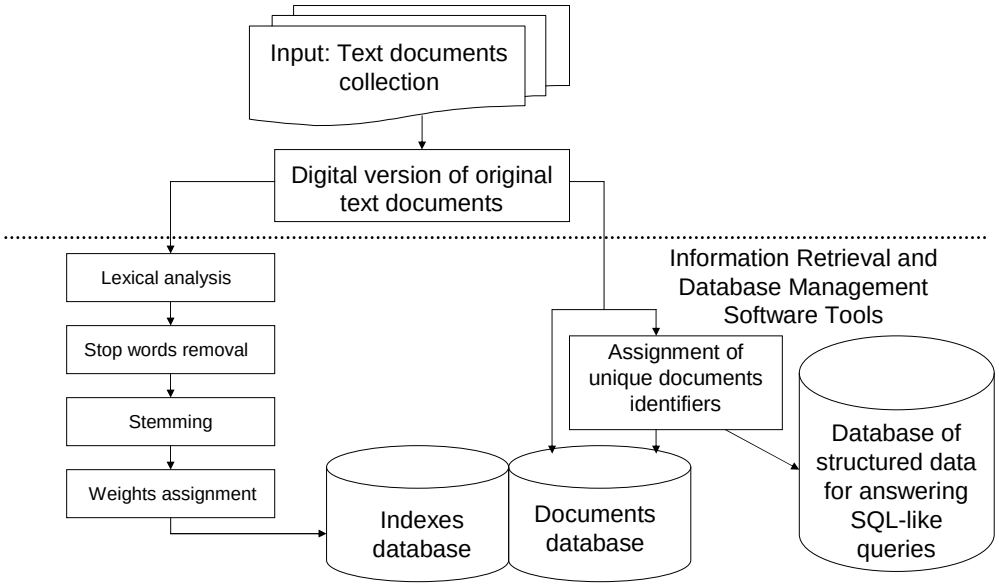
Document Gathering through a WSA



Indexer

Indexing the Web pages/documents

Indexing: Traditional IR



Indexing for Web Retrieval

Query-based Engines:
indexes are automatically built

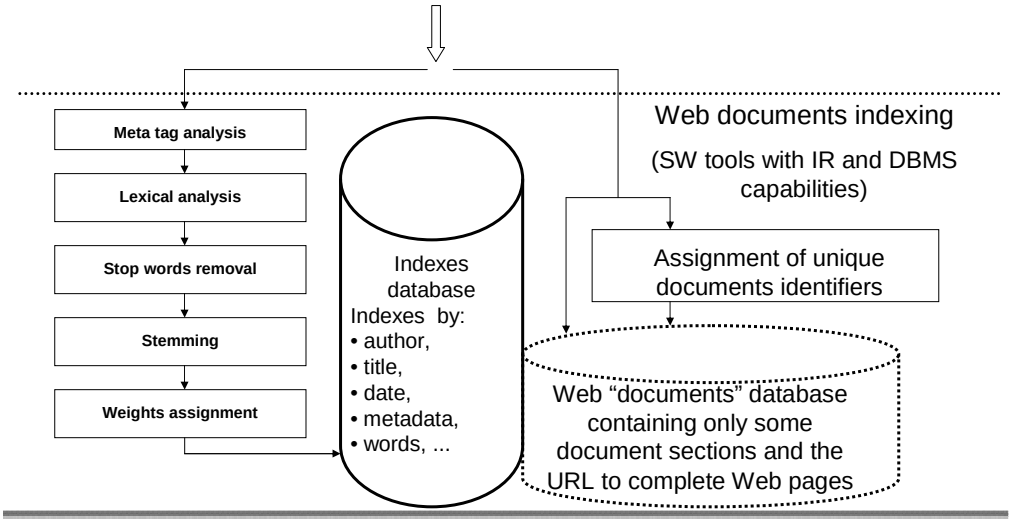
automatic

Classified Lists:
subject directory catalogues are manually built

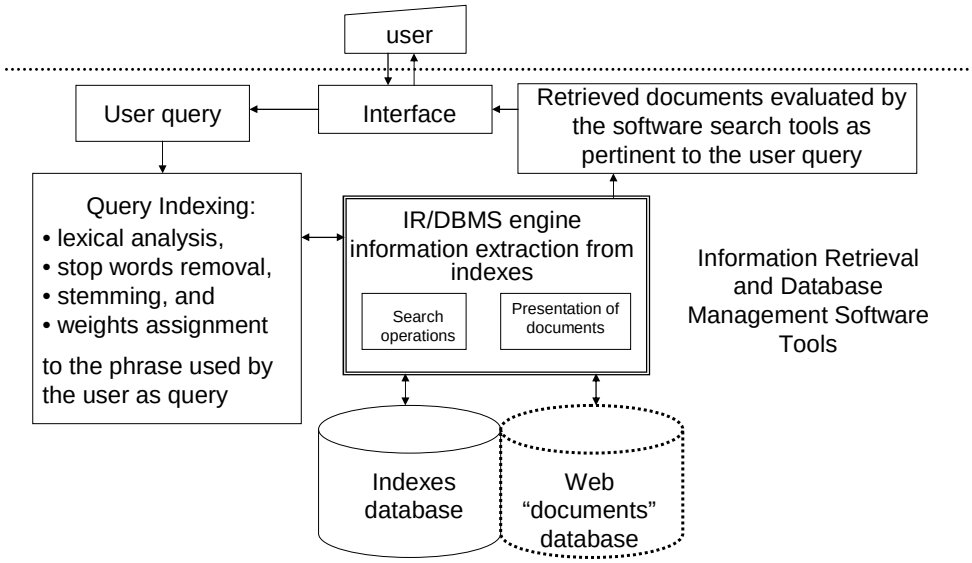
manual

Indexing for Web Retrieval

Web Search Agent (WSA)



Searching



Document and Query Management

"virtual collection" (docs discarded after indexing) ⟷ versus ⟷ "physical collection" (docs maintained after indexing)

The decision of maintaining the original version of Web documents that have been indexed can be made, because the document can change over time and the new one can be really different from the version which has been indexed, and it can constitute a surprise for the user when it sees it as present in the answer of its query.

Digital Libraries

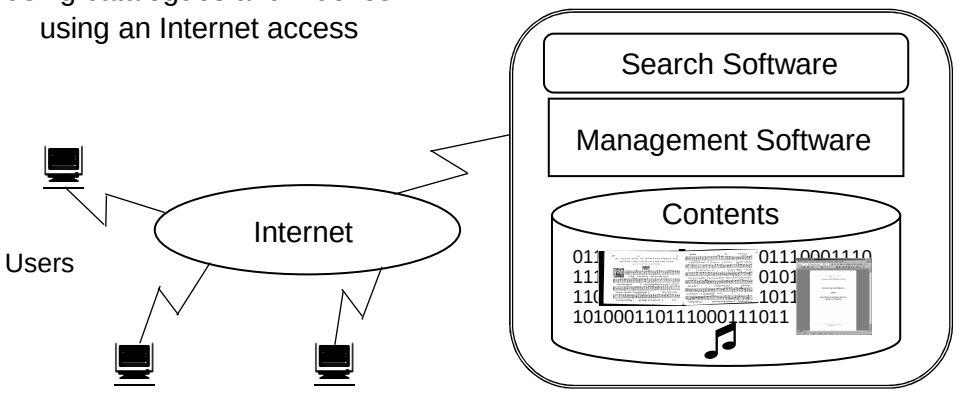
Search services usable through digital library architectures for the retrieval of digital documents

Digital Library: Internet Access

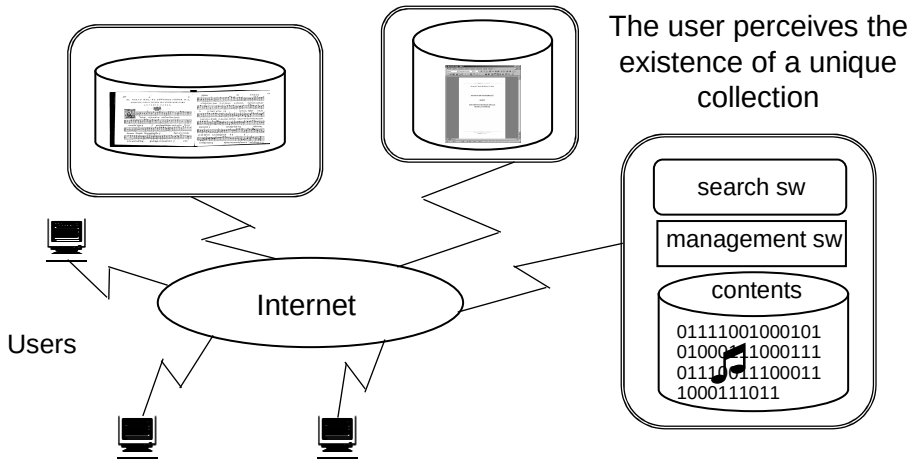
Search for documents

using catalogues and indexes using an Internet access

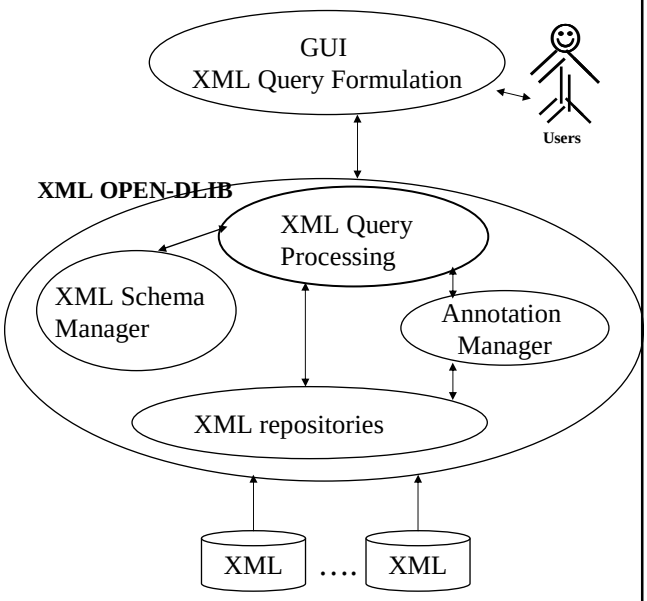
Digital library system



DL: Distributed Collections

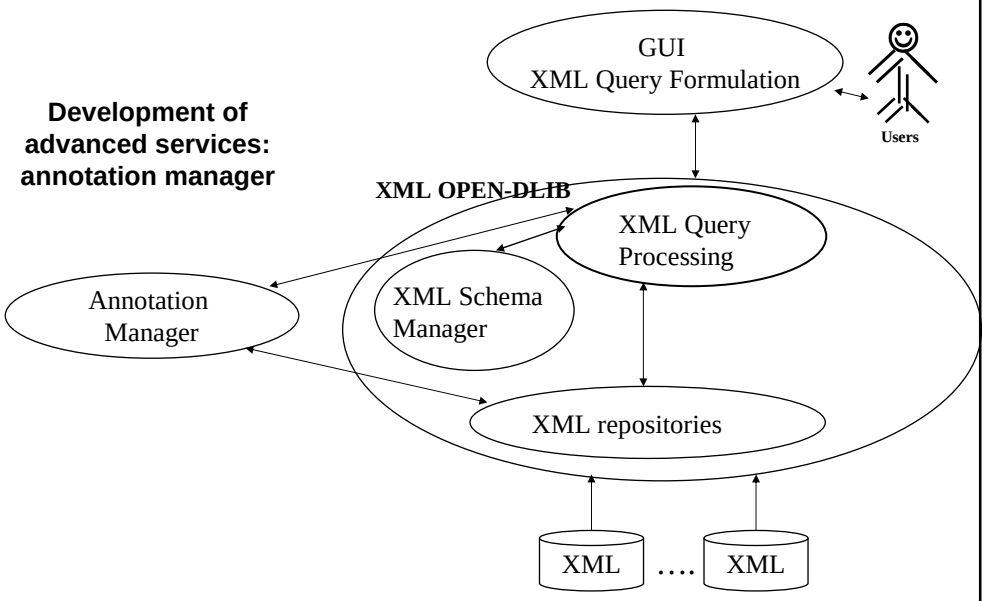


The Architecture of a DL



An Annotation Service in a DL Architecture

Development of advanced services: annotation manager

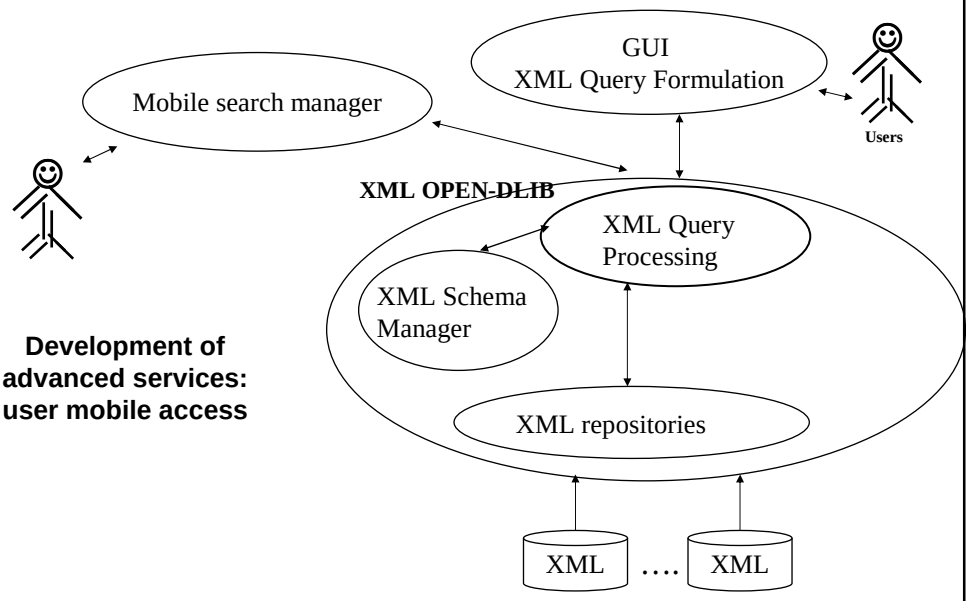


Annotation Service

Annotations:

- what:
  - references between documents or parts of documents
  - to highlight parts of text
  - to make notes
- where:
  - granularity: document, paragraph, section, phrase
  - annotations of annotations
- for the purpose of enriching the content of the DL

Extension to a DL Architecture



Development of advanced services: user mobile access

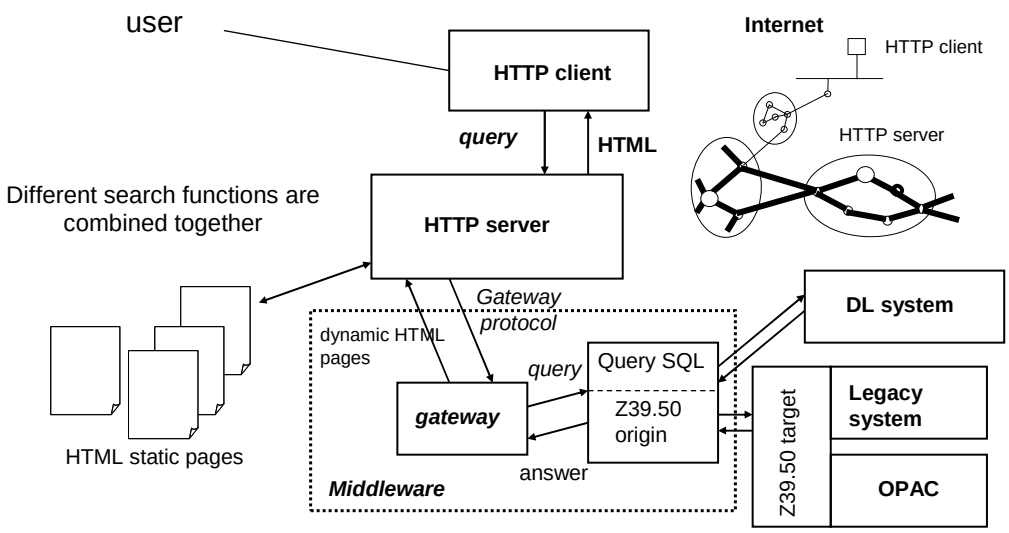
Catalogue Databases

Bibliographic/catalogue services for the retrieval of catalogue information and the localization of digital and traditional library material

Catalogue Databases

- Bibliographic and catalogue databases systems that provide services that are based on:
  - classification schemes that organize materials at a general level,
  - subject headings that provide a more detailed access, and
  - authority files that control variant versions of key information, such as:
    - geographic names, and
    - personal names

DL and Catalogue Databases: Contents Access



Suggested Bibliography on IR

- Information Retrieval:
  - [Porter, 1980] M.F. Porter. An algorithm for suffix stripping. *Program*, 14 (3), 1980, 130-137.
  - [Salton and McGill, 1983] G. Salton, M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY, USA, 1983.
  - [Sparck-Jones and Willett, 1997] K. Sparck-Jones, P. Willett. *Readings in Information Retrieval*. Morgan Kaufmann, San Francisco, CA, USA, 1997.
  - [van Rijsbergen, 1979] C.J. "Keith" van Rijsbergen. *Information Retrieval (2nd Ed).* Butterworths, London, UK, 1979.
    Online: http://www.dcs.gla.ac.uk/Keith/Preface.html.
  - [Witten et al, 1999] I.H. Witten, A. Moffat, T.C. Bell. Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann, San Francisco, CA, USA, 1999.

Suggested Bibliography on DL

- Digital Libraries:
  - [DELOS, 2001] DELOS. Digital Libraries: Future Research Directions for a European Research Programme. *DELOS Network of Excellence Brainstorming Workshop*, June 2001, San Cassiano (Dolomites), Italy.
    Online: http://delos-noe.iei.pi.cnr.it/activities/researchforum/Brainstorming/.
  - [Fox and Urs, 2002] E.A. Fox, S.R. Urs. Digital Libraries. In: B. Cronin (Ed), *Annual Review of Information Science and Technology (ARIST)*, Information Today, 36, 2002, 503-589.
  - [Witten and Bainbridge, 2003] I.H. Witten, D. Bainbridge. *How to Build a Digital Library.* Morgan Kaufmann, San Francisco, CA, USA, 2003.

Other Relevant References

- Information Agents:
  - [Klusch, 1999] M. Klusch (Ed). *Intelligent Information Agents: Agent-Based Information Discovery and Management on the Internet*. Springer-Verlag, Berlin, Germany, 1999.
- Authority Files:
  - [Hodge, 2000] G. Hodge. *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Council on Library and Information Resources (CLIR). April 2000.
    Online: http://www.clir.org/pubs/reports/pub91/.

Author's Addresses

- Author's Mail Address:
  - Department of Information Engineering
    University of Padova
    Via Gradenigo, 6/a
    35131 Padova, Italy
- Author's E-mail Address:
  - maristella.agosti@dei.unipd.it, or
    maristella.agosti@unipd.it
- Web page in the Research Group Web site:
  - Information Management Systems (IMS) Research Group
    IMS Web Page: http:/www.dei.unipd.it/~ims/