Poseidon House
Castle Park
Cambridge CB3 0RD
United Kingdom

TELEPHONE: **Cambridge (01223) 515010**
INTERNATIONAL: **+44 1223 515010**
FAX: **+44 1223 359779**
E-MAIL: **apm@ansa.co.uk**

ANSA Phase III

Metadata Presentation for CNET

Mark Madsen

Abstract

There is growing recognition of the need to leverage metadata (information about information) in order to automate the functioning of systems as well as facilitate the autonomous collection of information by agents.

This presentation briefly describes the status of available metadata research and indicates some of the areas in which it is expected to be a key component of next-generation systems.

APM.1772.01 **Approved** 30th May 1996

Briefing Note

Distribution:
Supersedes:
Superseded by:

Metadata: For a fruitful fishing in the information sea

Presentation for CNET, May 1996

Mark Madsen & Youcef Laribi

msm@ansa.co.uk & yl@ansa.co.uk

What is metadata ?

- Data that describes resources (individuals, organisations, software, documents, etc.).
- Examples:
  - SGML DTDs.
  - Relational Database schemas.
- Used for:
  - Learning about data (e.g Dictionary).
  - Grouping and linking related data sets (e.g Yellow Pages).
  - Discovering and locating data (e.g library indexes).

Gathering metadata.

- Extracted automatically by tools.
- Manually entered by the author, librarian, etc.
- Semi-automated (reviewed, updated or completed by humans).

Automatic extraction of metadata in the WWW

- The use of Web Crawlers, Worms, Spiders, Robots, Wanderers, etc.
  - WWWW (WWW Worm).
  - WebCrawler.
  - Lycos & Digital's Alta-Vista.
- Parses a full document (e.g Lycos) or parts of it (e.g WWWW).
- Creates searchable indexes out of the collected data.

Limitation of Web robots usage

- Create heavy network traffic.
- Intrusive (create excessive load on web sites) => Robot ethics protocol
- Build very large indexes (not scaleable).
- Imprecise results and limited query interface.
- Difficult to combine indexes produced by different robots.

Manual gathering of metadata

- Cataloging data into directories (e.g Yahoo).
- Advantages:
  - Allows for a detailed description of a resource.
  - Allows contextual searching (e.g under “Computing” Category, search “networks”).
  - Describes non textual information (images, audio, video).
  - Describes non-electronic resources (e.g individuals, organisations, etc.).
- Limitations:
  - Long and repetetive process.
  - Not cost effective.

Metadata formats

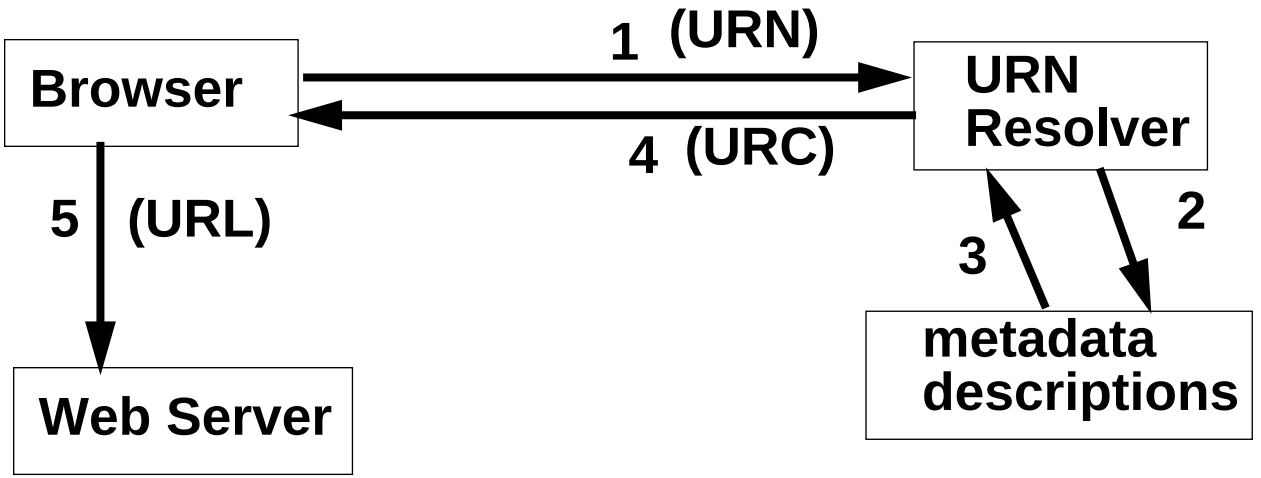
- **Several formats used in current information systems:**
  - **MARC (used by the librarian community).**
  - **Dublin Core (13 attributes. (e.g author, subject, date, ...)) for DLOs.**
  - **IETF IAFA WG defines template types and attributes per template type.**
  - **WHOIS++ centroids.**
  - **SOIF in Harvest.**
- **Suggestion of hierarchical/nested descriptions (simple/in-depth metadata).**

Deploying a metadata infrastructure

- Components for gathering, summarising, indexing, publishing and querying metadata.
- WHOIS++ directories and hierarchical index servers.
- Harvest (University of Colorado) comprises:
  - Summariser: produces SOIFs based on file types and administrator hints.
  - Gatherer: Stores SOIFs from one or more information sites.
  - Broker: Domain-specific directory: Collects SOIFs from one or more Brokers/Gatherers.
  - Search Engines: Indexes Broker's SOIF database and supports a query interface on the index.

Using metadata for better WWW naming schemes

- IETF URNs to uniquely identify resources independently of their physical location (e.g URN:ISBN:914256789x).
- URNs can be viewed as a query to retrieve location information and other metadata about the resource => URCs.



Using metadata for finding resources

- **Use search engines on metadata indexes, combined with query interfaces and languages:**
  - **Keyword search combined by boolean operators (e.g Glimpse).**
  - **Concept/Topic search (e.g Verity).**
  - **Attribute/Constraints search (e.g Infoseek).**
  - **Structured search (e.g SQL type queries).**
  - **Natural language querying.**

Metadata deployment recommendations

- Gather summaries as near as possible from the information source.
- Combine automatic and manual gathering.
- Support popular metadata formats (e.g SOIF, IAFA, URCs).
- Store summarised information (e.g SOIF) into standard directories (e.g LDAP/X.500 directories).
- Allow the support of more than one search engine/indexing technology (Brokers).
- Extensively use caching of metadata and even data if possible.
- Replicate directories and indexes for better performance and fault-tolerance.
- Allows for cascading directories/indexes for metadata refinement and better querying response times.