Metadata and Retrieval for Digital Archives

Prof. Ray R. Larson
University of California, Berkeley
School of Information Management and Systems
http://www.sims.berkeley.edu/
Overview

- XML and SGML for Digital Archives and Digital Libraries
- The Encoded Archival Description (EAD)
- The Cheshire II system and Features
- The Archives Hub and Distributed Archives Hub
Overview

• XML and SGML for Digital Archives and Digital Libraries
  • The Encoded Archival Description (EAD)
  • The Cheshire II system and Features
  • The Archives Hub and Distributed Archives Hub
What is SGML/XML?

• SGML stands for Standard Generalized Markup Language
  – XML stands for eXtended Markup Language

• What it is NOT:
  – Not a visual document description
  – Not an application specific markup
  – Not proprietary
What is SGML/XML?

• What it is:
  – An international standard (SGML- ISO 8879:1986)
  – A generic language for describing the structure of documents, and markup that can be used for those documents
  – Intended for generating markup for content rather than form elements

• XML is a simplified subset of SGML (being established by W3C)
Why XML is Revolutionary

• XML enables those developing WWW resources to preserve any “document type” or “database schema” when they publish on the Web.

• XML enables web services to send self-describing “protocol messages” that can be understood by programs, not just “by eye”.

• This information cannot be encoded in HTML.

• XML-encoded information is smart enough to support new classes of Web applications.

Adapted from Dr. Robert J Glushko
XML Enables New Web Applications

• Data interchange between Web clients
  – use Web for application integration without information loss (example: product information in supply chain, EDI)

• Moving processing from server to client
  – reduce network traffic and server load (example: download airline schedule, find best flights without “back-and-forth” thrashing)

Source Dr. Robert J Glushko
XML Enables New Web Applications

- Multiple client-side views of same data
  - expert and novice versions
  - manager and worker versions
  - localization (currency or measurement conversions)
- “Information push” from personalized applications
  - selecting information based on user preferences (example: custom news feed by matching article keywords against user profile)

Source Dr. Robert J Glushko
The First Generation Web

.. making information accessible through browsers

Source: Dr. Robert J Glushko
Airline Schedule

Flight Information
United Airlines #200
San Francisco
9:30 AM
Honolulu
12:30 PM
$368.50

Source Dr. Robert J Glushko
<Title> Airline Schedule </Title>

<Body>

<H2> Flight Information </H2>

<H3> United Airlines #200 </H3>

<UL>
    <LI> San Francisco </LI>
    <LI> 9:30 AM </LI>
    <LI> Honolulu </LI>
    <LI> 12:30 PM </LI>
    <LI> $368.50 </LI>
</UL>

Source Dr. Robert J Glushko
Next Generation Web

.. making information and services accessible to computers (and people)

Source Dr. Robert J Glushko
<TransportSchedule Type="Airline">
  <Segment Id="United Airlines #200">
    <Origin> San Francisco </Origin>
    <DepartTime> 9:30 AM </DepartTime>
    <Destination> Honolulu </Destination>
    <ArriveTime> 12:30 PM </ArriveTime>
    <Price Currency="USD"> 368.50 </Price>
  </Segment>
</TransportSchedule>

Source Dr. Robert J Glushko
XML is a Foundation for Interoperability

.. exchange information in an application and vendor neutral format

WEB  XML  EDI

Format based  API based

Source Dr. Robert J Glushko
XML and Digital Archives

- XML provides a platform for data interchange and for search and manipulation of described data.
- XML/SGML provides a standard framework for sharing and distribution of Archival information using the Encoded Archival Description and Encoded Archival Context markup languages.
Overview

• XML and SGML for Digital Archives and Digital Libraries

• The Encoded Archival Description (EAD)

• The Cheshire II system and Features

• The Archives Hub and Distributed Archives Hub
Components of Archival Description

- Description of records
- Context of creation: creators
- Functions and activities documented in records
- Dedicated descriptive semantics and structure for each component
- Components interrelated with one another
Records: EAD

- Encoded Archival Description
  - Society of American Archivists and Library of Congress
  - Used internationally
  - English, Spanish, Dutch, French, and Chinese
- 1998, 2002
What EAD Is

• An emerging encoding and structural standard for archival description
  – Data structure
  – Communication/interchange
  – Finding aid / archival description

• ISAD(G)
What EAD Is Not

- Content standard
- Data value standard
- Archival management system
Principals of Record Description

• Respect de fonds
  – Provenance
  – Original order

• Hierarchical and symmetrical

• Inheritance of description
The EAD DTD

• The EAD DTD is very complex and permits considerable flexibility in expressing the description and topics of the archival collection.
• The main parts are outlined on the following slides, but include:
  – A header, including basic descriptive info.
  – Optional frontmatter
  – The archival description
• We will describe only a few of the top-level tags
Major Sections and DTD Defs

• EAD
  – `<!ELEMENT ead (eadheader, frontmatter?, archdesc) >`

• EADHeader:
  – `<!ELEMENT eadheader (eadid, filedesc, profiledesc?, revisiondesc?) >`
  – FILEDESC
    • `<!ELEMENT filedesc (titlestmt, editionstmt?, publicationstmt?, seriesstmt?, notestmt?) >`
Major Sections and DTD Defs

• The Archival Description:
  – `<!ELEMENT archdesc (runner*, did, (admininfo | bioghist | controlaccess | odd | scopecontent | organization | arrangement | add | dsc | dao | daogrp | note')*)>`

• The Descriptive Identification
  – `<!ELEMENT did (head?, (abstract | physdesc | note | repository | origination | unitdate | unitid | unittitle | container | physloc | dao | daogrp)*)>`
<ead>
  <eadheader>Contains info on title, author, creation date. etc. of the finding aid</eadheader>
  <archdesc>Wrapper that holds all descriptive info such as title, date, extent, biographical sketch, scope and content, administrative and arrangement info, notes, subject headings, etc.

  <c01>Series level description e.g. “Correspondence”

    <c02>Sub-series level description e.g. “Family Correspondence”

      <c03>File level description e.g. “Letters to William C. Tyrrell”

    </c03></c02></c01></archdesc></ead>
<EAD>
<EADHEADER LANGENCODING = "ISO 639">
<EADID>
GB 0133 TAB
</EADID>
<FILEDESC>
<TITLESTMT>
<TITLEPROPER>
Tabley Muniments
</TITLEPROPER>
</TITLESTMT>
<PUBLICATIONSTMT>
<PUBLISHER>
John Rylands University Library of Manchester
</PUBLISHER>
</PUBLICATIONSTMT>
<ARCHDESC LEVEL = "FONDS" LANGMATERIAL = "English">
<DID>
<REPOSITORY>
University of Manchester, John Rylands University Library of Manchester
</REPOSITORY>
<UNITID ENCODINGANALOG = "ISADG3.1.1." COUNTRYCODE = "GB" REPOSITORYCODE = "0133">
GB 0133 TAB
</UNITID>
<UNITTITLE LABEL = "Title" ENCODINGANALOG = "ISADG3.1.2.">
Tabley Muniments
</UNITTITLE>
<UNITDATE LABEL = "Dates of Creation" ENCODINGANALOG = "ISADG3.1.3.">
19th century
</UNITDATE>
<PHYSDESC LABEL = "Extent" ENCODINGANALOG = "ISADG3.1.5.">
1.24 cu.m
</PHYSDESC>
<ORIGINATION LABEL = "Creator" ENCODINGANALOG = "ISADG3.2.1.">
Warren, family, of Tabley, Cheshire
</ORIGINATION>
</ARCHDESC>
</EAD>
Example EAD Record (Hub)

<biohist encodinganalog = "ISADG3.2.2.">
  <head>
    Administrative/Biographical History
  </head>
  <p>
    The poet John Byrne Leicester Warren, later 3rd and last Baron de Tabley, of Tabley near Knutsford, Cheshire, was born in 1835, the son of the 2nd Baron de Tabley (1811-1887), and his wife, Catherina. His mother was Italian, the daughter of the count de Soglio, and Warren spent much of his early childhood with her in Italy and Greece. He was educated at Eton and Christ Church, Oxford. At Oxford he published a volume of poetry. Originally he published under the pseudonyms George F. Preston (1859-1862) and William Lancaster (1863-1868), but latterly under his own name.
  </p>
  <p>
    His early verse included
  </p>
  <title>
    Praeterita
  </title>
  (1863),
  <title>
    Eclogues and Monodramas
  </title>
  (1864),
  <title>
    Studies in Verse
  </title>
  (1865),
  <title>
    Philocletes
  </title>
  (1866), and
  <title>
    Orestes
  </title>
  (1868). His early work was Tennysonian in style, but he was later to be influenced by both Browning and Swinburne. In 1873 he produced .... (some data removed)
Example EAD Record (Hub)

<SCOPECONTENT ENCODINGANALOG = "ISADG3.3.1.">
  
  <HEAD>
  Scope and Content
  </HEAD>
  
  <P>
  The collection consists mainly of the personal papers of the 3rd Baron de Tabley. The papers reflect his interests in literature, politics, botany and numismatics and include correspondence with numerous prominent later Victorian figures. Attention should also be drawn to de Tabley’s extensive and important collection of armorial bookplates.
  </P>
  
  <P>
  Correspondents include Sir Mountstuart Grant Duff, Edmund Gosse, Lord Houghton, A.C.Benson, and Robert Bridges. There are volumes of Tabley’s essays and verse, as well as a considerable number of notebooks and loose manuscripts of verse and other writings. There are various bundles and boxes relating to “Coins”, “Botany”, “Poetry”, “Literary”, and “Financial” and bookplates.
  </P>
  
  <P>
  Preliminary survey list.
  </P>

<ADD>
  
  <OTHERFINDAID ENCODINGANALOG = "ISADG3.4.6.">
  
  <P>
  There is correspondence with the 3rd Baron de Tabley among the Edward Freeman Papers, held at JRULM. The Library also has custody of the important Tabley Book Collection.
  </P>

<RELATEDMATERIAL ENCODINGANALOG = "ISADG3.5.3.">
  
  <P>
  The family and estate papers of the Leicester-Warren Family of Tabley are held by Cheshire Record Office. Some of these papers were originally in the custody of the John Rylands University Library of Manchester.
  </P>

<SEPARATEDMATERIAL>
  
  </ADD>
<CTRLACCESS>
<HEAD>
Index terms
</HEAD>

<GEOGNAME SOURCE = "NCARULES">
<EMPH ALTRENDER = "a">Tabley Inferior</EMPH>
<EMPH ALTRENDER = "a-">Cheshire SJ7378</EMPH>
</GEOGNAME>

<PERSNAME SOURCE = "NCARULES">
<EMPH ALTRENDER = "surname">Benson</EMPH>
<EMPH ALTRENDER = "forename">Arthur Christopher</EMPH>
<EMPH ALTRENDER = "dates">1862-1923</EMPH>
</PERSNAME>

<PERSNAME SOURCE = "NCARULES">
<EMPH ALTRENDER = "surname">Bridges</EMPH>
<EMPH ALTRENDER = "forename">Robert Seymour</EMPH>
<EMPH ALTRENDER = "dates">1844-1930</EMPH>
</PERSNAME>

<PERSNAME SOURCE = "NCARULES">
<EMPH ALTRENDER = "surname">Duff</EMPH>
<EMPH ALTRENDER = "title">Sir</EMPH>
<EMPH ALTRENDER = "forename">Mountstuart Elphinstone Grant</EMPH>
<EMPH ALTRENDER = "dates">1829-1906</EMPH>
<EMPH ALTRENDER = "epithet">Knight</EMPH>
</PERSNAME>

<PERSNAME SOURCE = "NCARULES">
<EMPH ALTRENDER = "surname">Gosse</EMPH>
<EMPH ALTRENDER = "title">Sir</EMPH>
<EMPH ALTRENDER = "forename">Edmund William</EMPH>
<EMPH ALTRENDER = "dates">1849-1928</EMPH>
<EMPH ALTRENDER = "epithet">Knight</EMPH>
</PERSNAME>

<PERSNAME SOURCE = "NCARULES">
<EMPH ALTRENDER = "surname">Milnes</EMPH>
<EMPH ALTRENDER = "forename">Richard Monckton</EMPH>
<EMPH ALTRENDER = "dates">1809-1885</EMPH>
<EMPH ALTRENDER = "epithet">1st Baron Houghton</EMPH>
</PERSNAME>

<SUBJECT SOURCE = "LCSH">
<EMPH ALTRENDER = "a">Bookplates</EMPH>
</SUBJECT>

<SUBJECT SOURCE = "LCSH">
<EMPH ALTRENDER = "a">Botany</EMPH>
</SUBJECT>

<SUBJECT SOURCE = "LCSH">
<EMPH ALTRENDER = "a">Numismatics</EMPH>
</SUBJECT>

<SUBJECT SOURCE = "LCSH">
<EMPH ALTRENDER = "a">Poetry</EMPH>
</SUBJECT>

<EMPH ALTRENDER = "a">Modern</EMPH>

<EMPH ALTRENDER = "y">19th century</EMPH>
</SUBJECT>
</ARCHDESC>
</EAD>
People: EAC

• EAD is now complemented by “EAC” or the “Encoded Archival Context”
• It is another XML-based standard for encoding descriptions of record creators: corporate bodies, families, and individuals
• It is being developed as part of an international effort with hopes of being able to share this information among archives having materials related to those entities
Authority Control

- Identifying creator entities
- Recording name or names used by and for them
- Rule-based heading or entry formation and control
Relations

- Creators
- Records
- Functions and activities
- Each relation qualified by place and time
- Records evidence of people acting in particular places and times
Characteristics and Events

• Person
  – Sex, education, address, competencies, activities, affiliations, awards …
  – Biography

• Corporate body
  – Type, mandate, location, legal status, assets, structure…
  – Administrative history

• Family
  – Assets and structure, activities, location, legal status…
  – Family history
Overview

• XML and SGML for Digital Archives and Digital Libraries
• The Encoded Archival Description (EAD)
• The Cheshire II system and Features
• The Archives Hub and Distributed Archives Hub
Overview of Cheshire II

- It supports SGML and XML with components and component indexes
- It is a client/server application
- Uses the Z39.50 Information Retrieval Protocol, support for SRW, OAI, SOAP, SDLIP also implemented
- Server supports a Relational Database Gateway
- Supports Boolean searching of all servers
- Supports probabilistic ranked retrieval in the Cheshire search engine as well as Boolean and proximity search
- Search engine supports "nearest neighbor" searches and relevance feedback
- GUI interface on X window displays and Windows NT
- WWW/CGI forms interface for DL, using combined client/server CGI scripting via WebCheshire
- Scriptable clients using Tcl and (new) Python
- Store SGML/XML as files or "Datastore" database
Cheshire II Searching

Local

Remote

Z39.50

Scanned Text

Images

Internet

Z39.50

Z39.50

Z39.50

Z39.50

Query Image 108019

Backword for this image
SGML/XML Support

- Underlying native format for all data is SGML or XML
- The DTD defines the database contents
- Full SGML/XML parsing
- SGML/XML Format Configuration Files define the database location and indexes
- Various format conversions and utilities available for Z39.50 support (MARC, GRS-1)
Distributed Search Tasks

• Resource Description
  – How to collect metadata about digital libraries and their collections or databases

• Resource Selection
  – How to select relevant digital library collections or databases from a large number of databases

• Distributed Search
  – How to perform parallel or sequential searching over the selected digital library databases
An Approach for Distributed Resource Discovery

- Distributed resource representation and discovery
  - New approach to building resource descriptions based on Z39.50
  - Instead of using broadcast search across resources we are using two Z39.50 Services
    - Identification of database metadata using Z39.50
    - Extraction of distributed indexes using Z39.50 SCAN
Z39.50 Explain

- Explain supports searches for
  - Server-Level metadata
    - Server Name
    - IP Addresses
    - Ports
  - Database-Level metadata
    - Database name
    - Search attributes (indexes and combinations)
  - Support metadata (record syntaxes, etc)
Z39.50 SCAN

• Originally intended to support Browsing
• Query for
  – Database
  – Attributes plus Term (i.e., index and start point)
  – Step Size
  – Number of terms to retrieve
  – Position in Response set
• Results
  – Number of terms returned
  – List of Terms and their frequency in the database (for the given attribute combination)
Z39.50 SCAN Results

Syntax: zscan indexname1 term stepsize number_of_terms pref_pos

% zscan title cat 1 20 1
{SCAN {Status 0}
{Terms 20}
{StepSize 1}
{Position 1}
{cat 27}
{cat-fight 1}
{catalan 19}
{catalogu 37}
{catalonia 8}
{catalyt 2}
{catania 1}
{cataract 1}
{catch 173}
{catch-all 3}
{catch-up 2} …

zscan topic cat 1 20 1
{SCAN {Status 0}
{Terms 20}
{StepSize 1}
{Position 1}
{cat 706}
{cat-and-mouse 19}
{cat-burglar 1}
{cat-carrying 1}
{cat-egory 1}
{cat-fight 1}
{cat-gut 1}
{cat-litter 1}
{cat-lovers 2}
{cat-pee 1}
{cat-run 1}
{cat-scanners 1} …
Resource Index Creation

• For all servers, or a topical subset…
  – Get Explain information
  – For each index
    • Use SCAN to extract terms and frequency
    • Add term + freq + source index + database metadata to the XML “Collection Document” for the resource
  – Planned extensions:
    • Post-Process indexes (especially Geo Names, etc) for special types of data
      – e.g. create “geographical coverage” indexes
MetaSearch Approach

- MetaSearch Server
- Internet
- Distributed Index
- Search Engine

Map Explain And Scan Queries
Map Results
Map Query
Map Results
Search Engine
Search Engine
Search Engine
Search Engine
Search Engine

DB 1
DB 2
DB 3
DB 4
DB 5
DB 6
Known Issues and Problems

• Not all Z39.50 Servers support SCAN or Explain

• Solutions that appear to work well:
  – Probing for attributes instead of explain (e.g. DC attributes or analogs)
  – We also support OAI and can extract OAI metadata for servers that support OAI
  – Query-based sampling (Callan)

• Collection Documents are static and need to be replaced when the associated collection changes
Our Collection Ranking Approach

- We attempt to estimate the probability of relevance for a given collection with respect to a query using the Logistic Regression method developed at Berkeley (W. Cooper, F. Gey, D. Dabney, A. Chen) with new algorithm for weight calculation at retrieval time.
- Estimates from multiple extracted indexes are combined to provide an overall ranking score for a given resource (i.e., fusion of multiple query results).
Titles only (short query)
Overview

• XML and SGML for Digital Archives and Digital Libraries
• The Encoded Archival Description (EAD)
• The Cheshire II system and Features
• The Archives Hub and Distributed Archives Hub
Distributed Archives Hub

• JISC-funded project in the UK (national service)
• Centralized Archives Hub at MIMAS (Manchester Computing)
  – Includes many HE sector archival collections totaling over 12,400 EAD descriptions
  – Using Cheshire II search software
  – Z39.50 cross-searching with National Archives
• Distributed set of Cheshire Servers with EAD-based archival collections
  – Currently still in development at Liverpool
  – Collections owners maintain their own servers
  – Uses the distributed search software of Cheshire II
Current Beta-Participants

• Spokes
  – University of Edinburgh, Murray Library
  – John Rylands University Library, Manchester
  – Leeds University, Brotherton Library
  – University of Liverpool
    • 30 virtual spokes representing the full Hub collection

• Full Contents of Central Archives Hub

• National Archives of Scotland and Wales will be participants
New Features

• Geographic search of archival collections
  – Using geographic interface
  – Searches mapped to placenames in search area using NIMA gazetteer
  – Ranked search over Geographic names

• Demo
  – Base Distributed Hub
  – UK_Geo prototype
Distributed Archives Hub Servers

- Search Servers
- Group Servers
- Replicated servers
- EAD Servers
Further Information

• Full Cheshire II client and server is open source and available at ftp://cheshire.berkeley.edu/pub/cheshire/
  – Includes HTML documentation
• Project Web Site http://cheshire.berkeley.edu/
• Cheshire III http://cheshire3.sourceforge.net/
• Archives Hub (centralized) http://www.archiveshub.ac.uk/