XML Element Retrieval and Heterogeneous Retrieval: In Pursuit of the Impossible?

Ray R. Larson
School of Information Management and Systems
University of California, Berkeley

Overview

- The Problem
- Issues with Element Retrieval and Heterogeneous Retrieval
- Possible Approaches
  - XPointer
  - Generic Metadata systems
    - E.g., Dublin Core
  - Other Metadata Systems

The Problem

- The Adhoc track in INEX has dealt with a single DTD for one type of data (computer science journal articles)
- In “real-world” environments, XML retrieval must deal with different DTDs, different genres of data and widely varying topical content

The Heterogeneous Track

- Research Questions (2004):
  - For content-oriented queries, what methods are possible for determining which elements contain reasonable answers? Are pure statistical methods appropriate, or are ontology-based approaches also helpful?
  - What methods can be used to map structural criteria onto other DTDs?
  - Should mappings focus on element names only, or also deal with element content or semantics?
  - What are appropriate evaluation criteria for heterogeneous collections?

INEX 2004 Het Collection Tags

Collection	Author tag	Title tag	Abstract tag
INEX (IEEE)	fm/au	fm/tig/atl	fm/abs
Berkeley	Fld100 Fld700	Fld245	Fld500 (*rarely*)
compuscienc e	author	title	abstract
bibdbpub	author altauthor	title	abstract
dblp	author editor	title booktitle	*None*
hcibib	author	title	abstract
qmulcspub	AUTHOR EDITOR	TITLE	ABSTRACT

Issues with Element Retrieval for Heterogeneous Retrieval

- Conceptual Issues (user view)
  - To actually specify structural elements for retrieval requires that the user know the structure of the items to be retrieved
  - As the number of DTDs or schemas increase this task becomes more complex for both specification and for understanding
  - For “real world” XML retrieval, specifying structure effectively requires omniscience on the part of the user
  - The collection itself must be specified in some way (can the user know all of the collections?)
  - Users of INEX can’t do correct specifications for even one DTD…

Issues with Element Retrieval for Heterogeneous Retrieval

- Practical Issues (programmers view)
  - Most of the same problems as the user view
  - As seen in an earlier papers today the system must provide an interface that the user can understand, but maps to the complexities of the DTD(s)
  - But, once again, as the number of DTDs or schemas increase this task becomes increasingly complex for the specification of the mappings
  - For “real world” XML retrieval, specifying structure effectively requires omniscience on the part of the programmer to provide exhaustive mappings of the document elements to be retrieved
    - As Roelof noted earlier today, this rapidly can become a system that has too many options for a user to understand or use

Postulate of Impotence

- In summation we might suggest another ``Postulate of Impotence'' like those suggested by Swanson
  - You can either have heterogeneous retrieval, or precise element specifications in queries, but you cannot have both simultaneously

Possible Approaches

- Generalized structure
  - Parent/child as in Xpath/Xpointer
  - What about flat structures? (like most collections in the Het track)
- Abstract query elements
  - Use semantic representations in queries rather than structural representations
    - E.g. “Title” instead of //fm/tig/atl
  - What semantic representations can/should be used?

XPointer

- Can specify collection-level identification
  - Basically a URN attached to an Xpath
- Can also specify various string-matching constraints on Xpath
- Might be useful in INEX Het Track for specifying relevance judgements
- But, it doesn’t address (or worsens) the larger problem of dealing with large numbers of heterogeneous structures

Abstract Data Elements

- The idea is to remove the requirement of *precise* and *explicit* specification of structural elements and replace them with *abstract* and *implied* specifications
- Used in other heterogeneous retrieval systems
  - Z39.50/SRW (attributesets and elementsets)
  - Dublin Core (limited set of elements for search or retrieval)

Dublin Core

- Simple metadata for describing internet resources
- For “Document-Like Objects”
- 15 Elements (in base DC)

Dublin Core Elements

- Title
- Creator
- Subject
- Description
- Publisher
- Other Contributors
- Date
- Resource Type
- Format
- Resource Identifier
- Source
- Language
- Relation
- Coverage
- Rights Management

Title

- Label: TITLE
- The name given to the resource by the CREATOR or PUBLISHER

Author or Creator

- Label: CREATOR
- The person(s) or organization(s) primarily responsible for the intellectual content of the resource. For example, authors in the case of written documents, artists, photographers, or illustrators in the case of visual resources.

Subject and Keywords

- Label: SUBJECT
- The topic of the resource, or keywords or phrases that describe the subject or content of the resource. The intent of the specification of this element is to promote the use of controlled vocabularies and keywords. This element might well include scheme-qualified classification data (for example, Library of Congress Classification Numbers or Dewey Decimal numbers) or scheme-qualified controlled vocabularies (such as Medical Subject Headings or Art and Architecture Thesaurus descriptors) as well.

Description

- Label: DESCRIPTION
- A textual description of the content of the resource, including abstracts in the case of document-like objects or content descriptions in the case of visual resources. Future metadata collections might well include computational content description (spectral analysis of a visual resource, for example) that may not be embeddable in current network systems. In such a case this field might contain a link to such a description rather than the description itself.

Publisher

- Label: PUBLISHER
- The entity responsible for making the resource available in its present form, such as a publisher, a university department, or a corporate entity. The intent of specifying this field is to identify the entity that provides access to the resource.

Other Contributors

- Label: CONTRIBUTORS
- Person(s) or organization(s) in addition to those specified in the CREATOR element who have made significant intellectual contributions to the resource but whose contribution is secondary to the individuals or entities specified in the CREATOR element (for example, editors, transcribers, illustrators, and convenors).

Date

- Label: DATE
- The date the resource was made available in its present form. The recommended best practice is an 8 digit number in the form YYYYMMDD as defined by ANSI X3.30-1985. In this scheme, the date element for the day this is written would be 19961203, or December 3, 1996. Many other schema are possible, but if used, they should be identified in an unambiguous manner.

Resource Type

- Label: RESOURCE TYPE
- The category of the resource, such as home page, novel, poem, working paper, preprint, technical report, essay, dictionary. It is expected that RESOURCE TYPE will be chosen from an enumerated list of types. One preliminary set of such types can be found at the following URL (now out of date):
  http://www.roads.lut.ac.uk/Metadata/DC-ObjectTypes.html

Format

- Label: FORMAT
- The data representation of the resource, such as text/html, ASCII, Postscript file, executable application, or JPEG image. The intent of specifying this element is to provide information necessary to allow people or machines to make decisions about the usability of the encoded data (what hardware and software might be required to display or execute it, for example). As with RESOURCE TYPE, FORMAT will be assigned from enumerated lists such as registered Internet Media Types (MIME types). In principal, formats can include physical media such as books, serials, or other non-electronic media.

Resource Identifier

- Label: IDENTIFIER
- String or number used to uniquely identify the resource. Examples for networked resources include URLs and URNs (when implemented). Other globally-unique identifiers,such as International Standard Book Numbers (ISBN) or other formal names would also be candidates for this element.

Source

- Label: SOURCE
- The work, either print or electronic, from which this resource is derived, if applicable. For example, an html encoding of a Shakespearean sonnet might identify the paper version of the sonnet from which the electronic version was transcribed.

Language

- Label: LANGUAGE
- Language(s) of the intellectual content of the resource. Where practical, the content of this field should coincide with the Z39.53 three character codes for written languages. See:

  http://www.sil.org/sgml/nisoLang3-1994.html

Relation

- Label: RELATION
- Relationship to other resources. The intent of specifying this element is to provide a means to express relationships among resources that have formal relationships to others, but exist as discrete resources themselves. For example, images in a document, chapters in a book, or items in a collection. A formal specification of RELATION is currently under development. Users and developers should understand that use of this element should be currently considered experimental.

Coverage

- Label: COVERAGE
- The spatial locations and temporal duration characteristic of the resource. Formal specification of COVERAGE is currently under development. Users and developers should understand that use of this element should be currently considered experimental.

Rights Management

- Label: RIGHTS
- The content of this element is intended to be a link (a URL or other suitable URI as appropriate) to a copyright notice, a rights-management statement, or perhaps a server that would provide such information in a dynamic way. The intent of specifying this field is to allow providers a means to associate terms and conditions or copyright statements with a resource or collection of resources. No assumptions should be made by users if such a field is empty or not present.

Issues in Dublin Core

- Lack of guidance on what to put into each element
- How to structure or organize at the element level?
- How to ensure consistency across descriptions for the same persons, places, things, etc.