World Library and Information Congress:
69th IFLA General Conference and Council

1-9 August 2003, Berlin

Code Number: 053-E
Meeting: 157. Bibliographic Control - Workshop
Simultaneous Interpretation: -

Metadata Schemas for Subject Gateways

Lynne C. Howarth
Faculty of Information Studies
University of Toronto
Toronto, CANADA
howarth@fis.utoronto.ca

Abstract:
Web-enabled subject gateways provide access to collections of high quality resources in a particular subject area. Descriptions of carefully selected and evaluated documents, objects, or sites uniquely identify and link to digital content. As this paper will examine, determining the focus and corresponding content for a subject gateway will necessarily influence the subsequent selection of metadata structures and schemas. Other considerations relating to metadata granularity and interoperability will also be assessed. The discussion will conclude with an analysis of challenges and future trends regarding the use of metadata schemas in subject gateways.

1.0 Introduction

As a simple search of Google ™ will confirm, operational definitions of the term, "subject gateways", are nearly as numerous as the services that they describe. Recognizing the need for a starting point, however, Traugott Koch (2000, 2) proposes the following as one approach to defining the concept:

> Subject gateways are Internet-services which support systematic resource discovery. They provide links to resources (documents, objects, sites of services), predominantly accessible via the Internet. The service is based on resource descriptions. Browsing access to the resources via a subject structure is an important feature.

Koch (2000, 1) notes further that, "Considerable manual effort is used to secure a selection of resources which meet quality criteria, and to display a rich description of these resources with standards-based metadata." With that additional explanation of

process, the connection between the concepts of "subject gateways" and "metadata schemas" is clarified, and the foundations laid for exploring how the latter can be used most effectively in the creation and support of the former. Particular challenges and future trends associated with metadata applications in subject gateways will also be discussed in later sections of the paper.

2.0 Considerations and Applications

In general, a distinction can be made between simple format metadata – such as that represented in the syntax of a mark-up language (e.g., XML; HTML; SGML), and embedded within the structure of the digital object – and structured rich format metadata. For the former, Web crawlers or "bots" can harvest the specified metatags (e.g., <Title>) to extract particular values, and, with minimal (or no) human intervention, arrange results within predefined inverted indexes or topic directories. The subject categories maintained at the sites for search engines Yahoo!™, AltaVista™, or Google™ are well-known examples. The problems that can arise with natural language vocabularies, unstructured text, and with what those in the bibliographic control community recognize as a lack of "authority control", are endemic in the "simple format" metadata environment. Yet the resistance to structure – either imposed or voluntarily adopted – in the free-wheeling Internet, continues to foster end-user initiatives, such as the Open Directory Project where resource creators can assign subject terminology that they have devised independently of any standard (see: http://www.dmoz.org)

In contrast, structured rich format metadata are devised, applied, and maintained in accordance with clearly established (international) standards. These are the formal metadata schemes that are used in "quality controlled" subject gateways (Koch 2000), and provide the value-add to resource discovery. In creating a "quality-controlled" metadata-enabled subject gateway, consideration must be given to three key aspects, namely, what resources (documents; objects; sites) to include, what kinds of metadata structures to use to describe and access those resources, and what metadata schema to apply in creating records to link to the content. Each of these aspects will be examined, in turn, in the sections that follow.

2.1 Determining Focus and Content for Subject Gateways

Determining the focus and corresponding content for a subject gateway will necessarily influence the subsequent selection of metadata *structures* and *schemas.*

Table 1
Inclusion Criteria for Determining the
Scope and Content of a Subject Gateway

Inclusion Criteria	Single (examples)	Multiple (examples)	Universal (examples)
Subject/Topic	Leukemia	Cardiac and Neurological diseases	All diseases
Language	English	German, French, Japanese, Greek	All languages
Geographic location	Canada	Europe and Asia	All countries, regions, etc.
Time period	2003	1900-1999	All recorded history
Type of resource	Web sites	Web sites, data repositories, and photo archives	All Web-enabled resources
Groups/Associations	Women	Children and young adults	Humankind
Format of material	Electronic text	Word documents, digital maps, DVD	All textual and media formats (analog and digital)

What is the specific intent of the subject gateway, what particular objectives are to be achieved in its design, and what deliverables or output are anticipated from it? The scope and coverage, delimitations and limitations of the subject gateway involve criteria, such as subject or topic, language, geographic location, time period, type of resource, groups or associations, format of material, etc. Table 1 summarizes, with examples, how each criteria could be combined and assessed to determine the final design of the subject gateway. Note that any number of approaches could be taken, and that the following offers only one such illustration where many permutations or combinations could be applied. The definition of "one", "many" or "all" of any of the criteria listed in the first column is also relative, and clearly open to interpretation.

Once these questions have been addressed, criteria for the selection of resources can be determined, and sources or targets for those resources identified. Koch (2000, 6) suggests that the most frequently occurring models for subject gateways include:

- National subject-specific (one subject; one country; one language – e.g., GEM)
- National cross-subject (multiple subjects; one country; one language – e.g., DutchESS)

- Global subject-specific (one subject; global; one language – e.g., EEVL)
- Global cross-subject (multiple subjects; global; one language – e.g., ADAM)
- Universal (all subjects; global; several languages – e.g., CORC)

One might add that the preceding typology is neither definitive nor exhaustive, and, like Table 1, adds one more view or perspective of the possible ways to combine selection criteria to design subject gateways based on their intended focus or specific objectives to be achieved.

2.2 Determining Applicable Metadata Structures

As noted previously, the determination of the scope, coverage, and selection criteria-driven content of a subject gateway will influence what metadata elements and schemas will be chosen to support identification of, and linkages to, targeted resources. In general, the *types* or *structures* of metadata that might be required to support a subject gateway modeled after any configuration of selection criteria outlined in Table 1, include the following:

- Administrative metadata: housekeeping" information about the record itself – its creation, modification, relationship to other records, etc. Examples of elements pertaining to administrative metadata include, but are not restricted to the following:
  - Record number
  - Date of record creation
  - Date of last modification
  - Identification of creator/reviser of record
  - Language of record
  - Notes
  - Relationship of this record to other(s)
- Descriptive metadata: describes the physical and intellectual properties or content of a digital item or object with such elements as:
  - Title (also alternative and parallel titles; subtitles; short titles; etc.)
  - Creator (author; composer; cartographer; artist; etc.)
  - Date
  - Publisher
  - Unique identifiers and dynamic links (URI; URL; etc.)
  - Summary; descriptive note; review; etc.
  - Audience level
  - Physical media; format; etc.
- Analytical metadata: information analysing and enhancing access to the resource's contents. Sometimes referred to as "subject metadata", elements may include:
  - Subject headings
  - Thesauri
  - Subject/topic keywords
  - Abstract; Table of Contents (TOC)
  - Classification codes derived from classification systems
  - Other elements of local importance, e.g., department affiliation; link to other related e-content; etc.

- Rights management metadata: information regarding restrictions (legal; financial; etc.) on access to, or use of, digital items or objects. Such elements as the following may apply:
    - Restrictions on use
    - Permission statements
    - Subscriber/licensing/pay-per-use fees
    - Acknowledgements
    - Copyright notice
    - Retention schedules
    - Quality ratings
    - Use disclaimers
- Technical metadata: particular hardware or software used in converting an item/object to a digital format, or in storing, displaying, etc., may require the use of such elements as:
    - Digitizing equipment specifications
    - Camera positions
    - Shooting conditions
    - Coding parameters
    - Voice recognition and/or read-back hardware and software
    - Optical scanner specifications
    - Image rendering equipment
    - Type of file and conversion software requirements
- Other, as determined – e.g., particular metadata elements based on local, regional, organizational requirements, or in accordance with a nationally mandated metadata standard, and not subsumed within any metadata type, above.

2.3 Selecting a Metadata Schema or Schemas

The choice of metadata schema or schemas to be used in creating the surrogate records for uniquely identifying and linking to resources accessible via the subject gateway will depend on the particular intent of the service and the types of metadata to be supported. Thus, a subject gateway, created and maintained by a distributed network of national organizations with content comprised of high quality Web sites (text and images, only), and limited to a subject area in a technical domain might require a mix of administrative, descriptive, and analytical metadata. The Canadian Health Network offers one example of such a configuration. A "virtual exhibit" containing links to a variety of digital objects contained within an international consortium of public and private art galleries and museums would necessitate the use of technical and rights management metadata, in addition to those required for administrative, descriptive, and analytical purposes.

What can help with the final determination of metadata schema is the desired degree of *granularity*, or, the amount of detail to be captured and represented in the metadata record. A "core record" – created using a metadata scheme, such as the Dublin Core with its fifteen element set (any of which are optional, repeatable, and extensible) – covers off adequately on administrative, descriptive, analytical, and rights management metadata, and can accommodate information related to technical specifications. In some specialized domains, however, a metadata schema, such as Dublin Core, lacks sufficient

granularity (detail) to adequately represent resources, or the particular purposes to which the subject gateway is directed. The ONIX metadata standard for international publishing and publishers, or the Content Standard for Digital Geospacial Metadata are two examples of rich, detailed, and highly technical metadata schemas, derived especially to deal with complex content and unique applications within the domain.

In addition to deciding on the level of detail to be captured in metadata-enabled records, the choice of schema can be narrowed in response to questions, such as the following:

- Is the proposed subject gateway in a (subject or discipline) domain for which a structured rich format metadata standard has been developed?
- Which fields would be most useful to the community of searchers the subject gateway is intended to service? How much detail should those fields support?
- Which fields would be most useful to those who are creating and/or maintaining the subject gateway? How much detail should those fields support?
- Which fields would be required to support particular services that the subject gateway is intended to provide?
- Will the use of, or access to, the subject gateway be restricted in any way? How will (should) this be recorded in the record metadata?
- Are there any requirements related to language, or format of material, or type of media for which particular (or additional) fields must be provided?
- Are there requirements to create or share resources among a network of collaborators with responsibility for the subject gateway? Are (additional) metadata fields required for gateway management?
- If the use of more than one metadata schema is envisioned or required (sharing resources across networks), are authoritative cross-schema mappings (crosswalks) readily and immediately available to facilitate and maintain interoperability? Can resources represented in one metadata schema (or standard) be exchanged with subject gateway collaborators who are using a different schema (or standard)?
- How widely used is a particular schema, and in what applications or environments comparable to the one currently proposed? How robust and/or flexible is the schema within different contexts?
- How readily can one migrate from this particular schema to another should data conversion be required at some time?
- How or how well does a particular schema comply with mandated organizational (local), national, or international standards, if any?
- What human (numbers; education; training), technical, financial, or other resources are required to support the application of the metadata schema, and does my organization or operation have those resources readily and sufficiently available? Are there other practical constraints to implementing and maintaining a particular schema or schemas?

Having answered any or several of the preceding questions, the choice to use one or more *standardized* metadata schemas may be confirmed. Alternatively, an individual, organization, or consortium electing to create a subject gateway may determine that a local or "home grown" solution – a set of locally-determined and supported metadata elements – is the preferred option. Similarly, some choose to combine elements of an

established standard, such as Dublin Core, with elements appropriate to the local situation of resources and objectives. There is no single recipe or "one-size-fits-all solution to which metadata schema or standard to use with a subject gateway.

3.0 Present Challenges and Future Trends

In general, present challenges are good predictors of issues that will require particular attention in the future, whether short- or longer term. As the number, coverage, scope, and end-user expectations of metadata-enabled subject gateways expand, a number of areas, such as the following, will be persistently problematic, and open to resolution:

- Interoperability – the requirement for enhanced cross-domain metadata protocols and crosswalks to support the exchange of records will grow; metadata standards to support interoperability at the technical, semantic, organizational, inter-community, and international levels may need to be developed or enhanced
- Collaboration and cooperation – subject gateways can be expanded using economies of scale to promote access to metadata-enabled access to cross-domain, international resources, as well as to share in the opportunities for, and costs of, creating and maintaining corresponding metadata schemas in common
- Scalability – While it is clear that some subject gateways are self-limiting, further growth in services designated as broad in scope or inclusion (e.g., OCLC's CORC; UKOLN) seems inevitable. While the realization of a truly universal subject gateway is unlikely in the very near future, it is a goal that should be anticipated, and gaps in metadata structures or elements addressed
- Multilingual resources – to-date metadata schemas have been developed and applied in monolingual environments; end-user demands for accessing multilingual resources in the language of their choice will require new or significantly expanded and enhanced metadata schemas, or innovative applications of existing non-verbal metadata schemas (e.g., classification systems) to describe and retrieve multilingual resources
- Search engine and interface functionality – increasingly sophisticated search engines will more effectively exploit whatever richness of metadata exists to support resource discovery; advanced cross-gateway searching or browsing may require the development of new or expanded schemas, such as collection description metadata, product or process metadata, metadata for new formats or technical innovations, etc.; likewise, new elements may be required to describe or support enhanced interface functionality (e.g., accessibility; usability; navigation; features to support special needs; etc.)
- Metadata toolkits – To-date, creating metadata records has been viewed as a largely manual (and human) exercise. As the number of subject gateways proliferates, automatic approaches will be increasingly applied, if only as a first intervention requiring subsequent human mediation. Tools that harvest and automatically index resources, populating pre-defined metadata record structures, will be used increasingly for subject gateway management
- Registries and local usage – refinements and expansions to existing metadata schemas has resulted in the creation of registries to record and track changes; the number of local variations to the standard have accelerated the growth in

registries; in future, there may be a movement to discourage or limit local, non-standard applications regardless of the availability of a registry

- Policies, legal issues, authentication – growth in subject gateways may necessitate expanding or enhancing metadata related to gateway management, intellectual property rights (IP), and the originality/authenticity of resources
- Standards compliance – ensuring, or even enforcing compliance with international, or cross-domain metadata standards or protocols may assume a priority as subject gateways proliferate and more collaborators are engaged

As the preceding may serve to illustrate, the range and diversity of issues arising from the application of metadata schemas in subject gateways are sufficiently numerous to engage metadata researchers and practitioners, alike, for some time to come.

Selected References:

Campbell, Debbie (2000). *Australian subject gateways: metadata as an agent of change*. Presented at the VALA 2000 Conference, Melbourne, Australia, 18 February, 2000. 7 pp. Available at: http://www.nla.gov.au/nla/staffpaper/dcampbell2.html Accessed 10/05/03.

DESIRE Information Gateways Handbook. 2000. 149 pp. Available at: http://www.desire.org/handbook/print4.html Accessed 30/04/03

De Jong, Annamieke (2002). Audiovisual domain. In *SCHEMAS Metadata Watch Report #8 and Standards Framework Report #4: Appendix A*. 6 pp. Available at: http://www.schemas-forum.org/metadata-watch/d29/d29.htm Accessed 24/4/03

Howarth, Lynne C. (2003). *Metadata unplugged*. A presentation to the 25th Anniversary Conference of Substance Abuse Librarians and Information Specialists, Toronto, Canada, 25 April, 2003. Available at: http://www.SALIS.org/ Accessed 19/05/03

Koch, Traugott (2000). Quality-controlled subject gateways: definition, typologies, empirical overview. *Online Information Review*. Vol. 24:1. 17 pp. Available at: http://www.lub.lu.se/tk/publ/OIR-SBIG.html Accessed 07/05/03.