Towards the "webification" of controlled subject vocabulary

A case study involving the Dewey Decimal Classification

Michael Panzer

Global Product Manager, Taxonomy Serivces

OCLC

6th European NKOS Workshop

September 21, 2007

Budapest

Introduction/Anamnesis

- **Well-known problem, why does it resurface every other year without much change?**

- **Large scale projects dealing with KO vocabularies are started without adhering to common fundamentals on the operational and strategic level**

- **Project results are often unsustainable and do not outlive the specific use case (if any) that they were build to support**

- **Currently, the DDC is facing such a challenge and chance for transition to the "network level"**

  - "Network level": Infrastructural improvements to make a KOS web-scale accessible, to make sharing, syndicating, leveraging of its data feasible

- **Main project goal: Improving accessibility and visibility of the scheme to stimulate association with resources**

A. Restating the Obvious: Some Truisms of Structural and Infrastructural Improvement

1. **Design of identifiers**

2. **Design of verbal designators ("verbal plane")**

3. **Data representation**

4. **Enhancement of the scheme itself**

5. **User contribution**

6. **Versioning**

7. **Vocabulary registries**

A. Restating the Obvious: Some Truisms of Structural and Infrastructural Improvement

1. **Design of identifiers**

2. **Design of verbal designators ("verbal plane")**

3. **Data representation**

4. **Enhancement of the scheme itself**

5. **User contribution**

6. **Versioning**

7. **Vocabulary registries**

1. Identification

- **Addressability and reference as problem for the web at large**

- **Rigorous semantic engineering ("subject landscaping," G. Dunsire) in KOS often not fronted for outside use**

- **Landscaping becomes just baroque gardening, withdrawing the horticultured space from the landscape at large**

2. Verbal designation

- **Scant research about what to show end-users and how to do it**

- **Primarily treated not as a question of semantics, but usability**

- **Usability usually not an attribute of terminologies, but their user-facing services (front-ends)**

- **Starting from scratch not possible, and transformation not trivial**

  - intricate interdependencies and contextual configurations of infons/taxons in the KO systems

3. Data representation

Different levels of disarray, different aggregate states:

- **Printed form, proprietary formats, spreadsheets**

- **Accessibility limited to specific communities**

- **But: emerging standards (SKOS, OWL-DL) are often under consideration for adoption**

- **Again: crosswalking is all but trivial; sometimes conceptual properties of the KOS have to be adapted**

B. Webifying the DDC (the first steps)

- **URI design**

- **Caption design**

- **Format considerations**

It's the URI, stupid!

Premise:

- **To summon a demon, you need to know its name (or vice versa: if you want to be summoned, you should try to get your name out there)**

Importance of URIs:

- **Easy to remember**

- **Easy to share**

- **(Relatively) easy to compare**

- **Best practice formats (RDF, SKOS) are URI centric**

URIs for the DDC

Design goals:

- **Common locator for Dewey concepts and associated resources for use in web services and web applications**

- **Use-case-driven, but outlasting and not directly related to a specific use case (persistency)**

- **Retraceable path to a concept rather than an abstract identification, reusing a means of identification that is already present in the DDC and available in existing metadata**

URIs: Basic Format

http://dewey.info/{aspect}/{object}/{locale}/{type}/{version}/{resource}

{aspect} is the aspect associated with an {object}—the current value set of aspect contains "concept", "scheme", and "index"; additional ones are under exploration

{object} is a type of {aspect}

{locale} identifies a Dewey translation

{type} identifies a Dewey edition type and contains, at a minimum, the values "edn" for the full edition or "abr" for the abridged edition

{version} identifies a Dewey edition version

{resource} identifies a resource associated with an {object} in the context of {locale}, {type}, and {version}

URIs: Examples

<http://dewey.info/concept/338.4/en/edn/22/>

<http://dewey.info/concept/333.7-333.9/>

<http://dewey.info/concept/2--74-2—79/>

<http://dewey.info/concept/333.7-333.9:1:16/>

<http://dewey.info/scheme/en/edn/22/>

<http://dewey.info/index/African National Congress/en/edn/22/>

<http://dewey.info/concept/333.7-333.9/about.skos>

URIs: Open Issues

- **Order of DDC entities, placement of {locale} component**

  - What makes sense

    - from a data model standpoint?

    - from a services standpoint?

- **Identification of other Dewey entities**

  - External summaries, tables as a whole, different types of editions, optional numbers, DDC Manual

URIs: Further Considerations

- **Multiple URI schemes for different service contexts (if unambigous and compliant with httpRange-14)**

- **Different syntax specifications (EBNF vs. URI Templates)**

- **Opacity vs. traceability**

- **Risk of defining identifiers without a service**
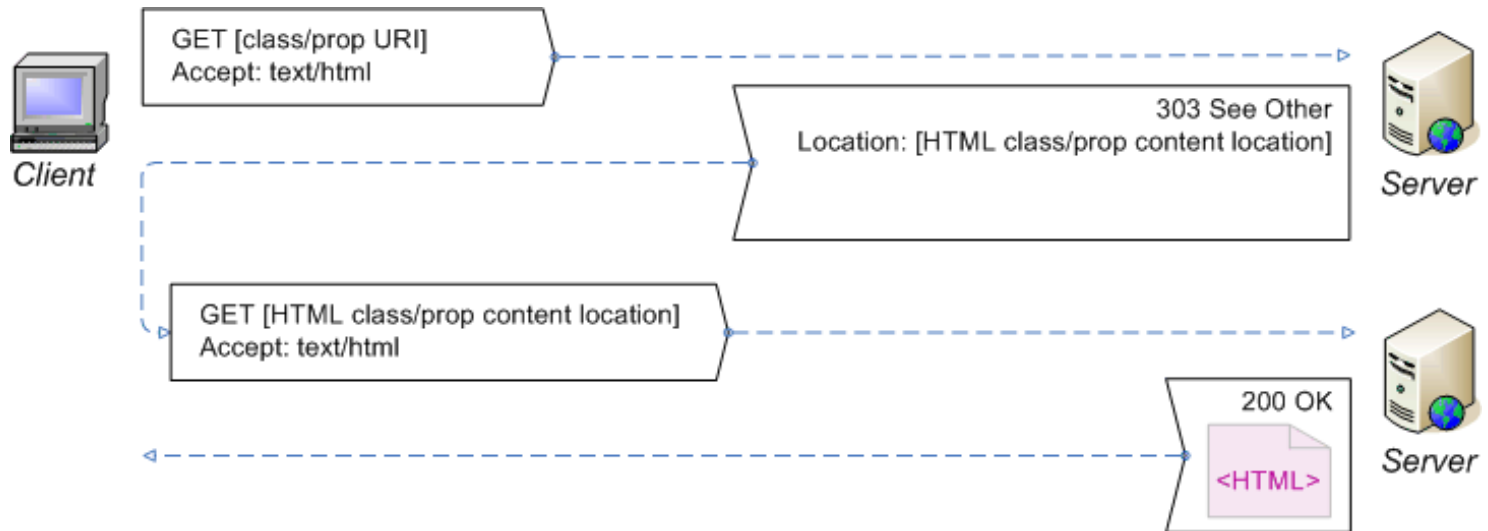
- **Location vs. identification as ontologial problem**

URIs: Layering Schemes

http://dewey.info/338.4/en   `Accept: text/html`

Server response: `303 See Other`

New location: http://dewey.info/concept/338.4/en/edn/22/about.html

Caption design

- **Problems specific to schemes depending on hierarchy**

- **Display context has to be taken into account**

- **Two different modes of display**

  - First level: Optimized for glanceability

  - Second level: Aggregated information from different sources

- **Process has to be at least "Pareto-automatic" (80/20)**

- **Improving captions by aggregation and mining of own and associated data**

Caption Design: Fundamentals

- **Comprehensibility of Dewey class headings highly dependent on the context of presentation**

- **Context is fixed in existing web applications, fluid (unknown) for web services**

- **Prediction of necessary information to give good impression of scope and meaning is not trivial**

Caption Design: Hierarchy I

- **Dependence on hierarchy to indicate discipline**

- **Folding parts of an hierarchical array back into the caption**

- **Smooshing the context to become useful for enriching the caption, without either flattening it completely or displaying it entirely**

- **Avoiding the drawbacks of a classic breadcrumbs display**

Caption Design: Hierarchy II

025.349 [Cataloging, Classification, Indexing of] Other special materials

- **Framing by discipline derived from the Relative Index:**

025.349 <u>Cataloging of o</u>ther special materials

- **Other strategies to acquire relevant contextual terminology:**

  - Relationship types in hierarchical array

  - Associated resources (and co-occurring subject vocabulary)

  - Mapped vocabulary

Caption Design: Heading types

- **"Other" headings**

- **Node labels**

  - Hook numbers

  - Centered entries

- **Brief headings**

- **'Deweyisms'**

- **Homonymity/polysemy in headings**

- **Standard subdivisions and other technical vocabulary**

Format considerations I: MARC 21

- **DDC migrating from proprietary format to MARC 21 Classification and Authorities (http://www.loc.gov/marc/marbi/2007/2007-dp06.html)**

  - Revamping of 082 field for better subject access (provisions for assigning internal table notation, external table notation, identification of standard/optional numbers)

  - Provision for additional Dewey numbers as access numbers

  - Inclusion of component parts of numbers in bibliographic records using a new 085 field

  - Identification of notation in internal add tables and (where not already provided) in Tables 1–6

- **MARC is at the epicenter of OCLC expertise**

- **Starting/transition point for a variety of crosswalks**

Format considerations I: MARC 21

- **Component Parts Example:**
  Feminist Criticism of Television

082 01 $8 1 $a 791.45082 $2 22

085 ## $8 1.1 $b <u>791.45</u> $z 1 $s <u>082</u> $u 791.45082

Television   **Feminist**

Format considerations II: SKOS

Feasibility of providing a SKOS version of Dewey:

- **Solving of the identifier issue**

- **Minor standard issues: collections, note types**

- **Concept versioning**

- **Representing the Relative Index**

- **Revitalizing SKOS Mapping Vocabulary Spec**

Thanks for participating!

Questions, comments, discussion:

Michael A. Panzer, panzerm@oclc.org