Metadata Aggregator Task Force

Some “Findings-To-Date”
Task Force Charge

high-level functional requirements...

- the list of candidate metadata sources (e.g. Barton, Vera, DSpace, GIS, VDC, IRIS, archival findings aids, ...)
- the list of source metadata schemas (e.g. MARC, qualified DC, FGCD, DDI, VRA, EAD, or custom schemas...)
- recommendations regarding metadata ... normalization in the tool for ingest, faceting, indexing, and display
- what facets will be initially generated, and whether they must be the same for all metadata sources in the tool
- any search requirements beyond keyword search of all metadata...
- any additional features of the tool or interface, such as user tagging, rating, reviews or other annotation...
Two “information problems”

• Too many places to start

• Too many results
Simplified Search

- Too many places to start from
- More like Google, pls.

MIT Libraries

I. Single Box

Search
Simplified Search

- Full-text type of search mindset...

MIT Libraries

hurricane katrina

Search
“Mixed” Results ;^)
with... Faceted Browse

- Facets expose collections
- Facets guide result sets
- Facets need some care & tending...

MIT Libraries

hurricane katrina

- Book 1 ...
- Book 2 ...
- Article 3 ...
- Book 4 ...
- Website 5 ...
- Book 6 ...
- Image 7 ...

Facet 1
- value 1a...
- value 1b...

Facet 2
- value 2a...
- value 2b...
Behind the Scenes

• Not technically
• Metadata-wise...
Libraries’ Content

- To be aggregated
- Some metadata, some ‘data’
Content’s Schemas...

- Content sources may have differing schemas
- (Within schemas, may have differing (or no!) controlled vocabularies...)
Facets (U/I)

- Selection of Facets = utility to end-user

Facets

- AUTHOR
- KEYWORD
- DATE
- ...

Metadata Schemas

- MARC
- DC
- VRA
- HTML
Mapping to Facets

- Many elements, even from one schema, may map to one facet
• Some elements may get processing on contents (e.g. parsing subject heads, reversing names, etc.)
Mapping to Facets (3)

- Some elements will require design decisions re: inclusion, format, normalization, conditional logic, concatenation, aggregation, code resolution etc.
Elsewhere...

California Digital Library
Melvyl “Recommender”

“We had to invest significant effort to work with bibliographic records: deciding what elements and attributes to index, investigating which MARC fields contained essential data, and implementing code expansions.”
Inverted Index (Full-Text)

- (Almost) All available text will be indexed
- Includes text not mapped to facets

- (Orange) = Full-text (.PDF, HTML)
- (Black) = Metadata elements

Facets

AUTHOR

KEYWORD

DATE

...

Metadata Schemas

MARC

DC

VRA

HTML

Index

EXTRACTED TEXT
Index Relation to Facets

- Non-scientific slide (!)
- Point: Faceted text may receive different treatment, processing, from full-text.
Facets by Metadata Schemas

- AUTHOR
- MARC
- DC
- VRA...

- KEYWORD
- MARC
- DC
- VRA...
Metadata Schemas by Facets

- **MARC**
- **AUTHOR**
- **KEYWORD**
- **DATE...**

- **DUBLIN CORE**
- **AUTHOR**
- **KEYWORD**
- **DATE...**
Metadata Schema (VRA-Core) by Facets

- **VRA-CORE**
- **AUTHOR**
- **KEYWORD**
- **DATE**
- **FORMAT**
- **TYPE**
- **LANGUAGE**
- **MATERIAL**
Facet (Keyword) by Metadata Schemas

- **KEYWORD**
- **MARC**
  - Title-related
  - Subject-related
  - Other text
- **DUBLIN CORE**
  - Title-related
  - Subject-related
  - Other text
- **VRA-CORE** ...
- **HTML**
Simplified Search (redux)

- Single search supported by aggregated index of text from all content sources
Faceted Browse (redux)

User Interface

MIT Libraries

Book 1 ...
Book 2 ...
Article 3 ...

Facet 1

Back End

Facets

AUTHOR

KEYWORD

DATE

Facets Index

FACETED TEXT

Index

EXTRACTED TEXT

- Faceted browse supported by selection and mapping of metadata elements from all content sources
Supplementary Metadata...

These are Linked or Matched to “Records”

- **Access / Licensing** - ERM, SFX, Vera/Verde, KB
- **Availability** - current circ status
- **Popularity** - circ data; emetrics, SFX
- **Topical Access** - Ulrich’s; other
- **Evaluative** - reviews, book covers...
- **User-Generated Content** - tags, comments...
“Meta”-Metadata...

- Content about content about content...
- LCSH - *not* attached to any record. The hierarchy per se, as aid to further search
- Future possible sources...
  - Wikipedia + Yahoo terms
  - del.icio.us tags
  - LibraryThing
Task Force Charge

...for acquiring a Metadata Aggregator

- Identify **candidate products** available in the current marketplace and select three for detailed review.

- Develop and implement a **product review process** to be approved by the project’s steering committee for the three candidate products.
Marketplace of “Solutions”

Commercial
- Endeca
- Siderean Seamark
- Primo (Ex Libris)
- Dieselpoint
- i411
- FAST

Open Source
- MIT, Simile Longwell, Dwell
- Apache Solr (Lucene)
- Berkeley, Flamenco
- U. Rochester, ‘XC’ - Extensible Catalog
Vendor Considerations

• Meeting requirements (metadata, facets, U/I, trusted network features, others...)
• Scalability, performance
• Development tools, support, level of effort
• Total Cost of Ownership
• Timing of product release, maturity
• System-to-system processes (batch load, utilities, etc.)
• “Future-proofing” data against lock-in
Metadata Aggregator Task Force

Report due to Steering Committee within November