Got Data, now what?

Trish Rose  
Univ of CA, San Diego  
DLF Fall 2005  
Charlottesville, VA
UCAI

Goal: to develop a shared cataloging prototype for the visual resource community in order to promote the copy cataloging of image metadata

It was a Multi-phase project:
- Phase 1 (April 2002 - December 2003)
- Phase 2 (January 2004 - October 2005)
**UCAI processes**

- **Mapping** – to a single element set (VRA Core 3.0 extended)
- **Ingesting** – based on maps with some normalization
- **Clustering** – to bring records for the same work together
- **Merging** – for ease of viewing and navigating
Mapping and Ingest

- Data Submission
- Minimal record requirements
- Normalization
Mapping and Ingest
Data Submission Guidelines

Data Format
- XML or tab/comma delimited files
- character encoding (e.g. Unicode or Latin 1)
- external files (e.g. DTDs or Schemas)

Documentation
- data dictionary
- cataloging guidelines
- identifiers

Contacts
Mapping and Ingest

Minimal record requirements

- Title
- Agent or Cultural Group
- Date
- Site or Repository
- Work Type
Mapping and Ingest
Minimal record requirements

- Title
- Agent or Cultural Group
- Date
- Site or Repository
- Work Type
Mapping and Ingest
Minimal record requirements

- Title
- Agent or Cultural Group
- Date
- Site or Repository
- Work Type

Only 45% had all 5 populated
Mapping and Ingest

Minimal record requirements

- Title
- Agent or Cultural Group
- Date
- Site or Repository
- Work Type

Only 69% had 3 populated
Mapping and Ingest
Minimal record requirements

RLG’s levels
- base-line (should be included)
- value-added (provides enhanced functionality)
- bonus (contributes to optimal functionality)

from Descriptive Metadata Guidelines for RLG Cultural Materials
Mapping and Ingest Normalization

Establish plan

- Normalization types
- Normalization priorities
- Identify fields to be normalized
- Assign normalization responsibilities (provider or aggregator)
Mapping and Ingest Normalization

Types

- Parsing (no 1:1 relationship)
- Preferred term (reducing variants)
- Contextual data (e.g. brackets, codes, initials)
- Misspellings/factual errors (e.g. Picassi)
- Reformatting (e.g. late 18\textsuperscript{th} century, 1775-1800)
Factors to consider

when choosing which normalization to perform
   *Will it require automated or manual techniques?*

when choosing fields to normalize
   *Which fields will benefit users?*

when assigning normalization responsibilities
   *Does the provider or aggregator benefit?*