The Dryad Data Repository: A Singapore Framework Metadata Architecture in a DSpace Environment

Hollie C. White
Sarah Carrier
Jane Greenberg
Abbey Thompson
Ryan Scherle
Overview

• Recent metadata developments for Dryad (formerly known as DRIADE)
  – A digital data repository for datasets underlying publications in the field of evolutionary biology

• Implementation of the system in a DSpace environment
  – Current efforts to represent of the Dryad application profile in DSpace

• Bringing the Dryad application profile into conformance with the Singapore Framework
  – Challenges, considerations for the future
The Dryad Repository

• Dryad’s role and functionalities
• Collaboration between the SILS Metadata Research Center (UNC Chapel Hill) and the National Evolutionary Synthesis Center (NESCent)
• Two goals for metadata activities:
  – Dryad’s need to be interoperable with other data repositories used by evolutionary biologists
  – Dryad’s need for a sustainable information infrastructure
DSpace Implementation

• Benefits:
  – Adaptable, will support Dublin Core metadata
  – Submission system

• Challenges:
  – Modifications are difficult
  – Default workflow for submitting content is too cumbersome for users
  – Permissions issues
  – Metadata with hierarchical information (e.g. MODS) not supported in core repository

DSpace : http://www.dspace.org/
Dryad Application Profile, version 1.0

- Modular design
  - Data Object module
  - Publication module
- Incorporates elements from:
  - Dublin Core, Darwin Core, PREMIS, DDI, EML
- Supports Dryad functionalities
  - Basic data/metadata storage
  - Simple retrieval and submission system

Singapore Framework Compliance

- Standard for Dublin Core application profiles
- Benefits
  - consistency, long-term quality control, and interoperability with other metadata structures
- Use of Scholarly Works Application Profile (SWAP) as a key example of an application profile in conformance with the Singapore Framework
Functional Requirements

• Scope
• Stakeholders and designated community
  – Researchers in the field of evolutionary biology, publishers of established biology journals
• Requirements gathering
  – Stakeholders workshop, use case study, survey
• Functional requirements
  – Resource discovery and use
  – Data interoperability
  – Automatic and semi-automatic metadata generation
  – Linking of publications and underlying datasets
  – Data/metadata quality control
  – Data security
Domain Model

- Dryad application profile version 1.0 accommodates one publication associated with multiple datasets
Description Set Profile and Usage Guidelines

• DSP is “an information model and XML expression”
  – http://www.unc.edu/~scarrier/dryad/DSPLevelOneAppProfDraft.xml

• Usage guidelines are optional

Challenges and Future Work

• Ongoing revision of the Dryad application profile
• Streamlining Dryad’s interface for entering metadata
• Limitations in the current state of citation metadata
• Determine how or if elements from non-Dublin Core namespaces should be included in the DSP
  – Issues with interoperability, e.g. RDF/DCAM
• **Dryad**
  - [http://datadryad.org/](http://datadryad.org/)
  - **Dryad Wiki**
    - [https://www.nescent.org/wg_digitaldata/Main_Page](https://www.nescent.org/wg_digitaldata/Main_Page)
    - Includes links to publications, the application profile, and lists Dryad team members

• **Metadata Research Center <MRC>**
  - [http://www.ils.unc.edu/mrc/](http://www.ils.unc.edu/mrc/)

• **National Evolutionary Synthesis Center (NESCent)**

**Contacts:** Hollie C. White (hcwhite1@email.unc.edu), Jane Greenberg (janeg@email.unc.edu), Sarah Carrier (scarrier@email.unc.edu)
Dublin Core for datasets: DISC-UK approach

Robin Rice
EDINA and Data Library
University of Edinburgh
Scotland, UK
DISC-UK DataShare (Mar 2007-Apr 2009)

- A project led by EDINA at the University of Edinburgh, funded by JISC with partners the Universities of Oxford and Southampton (LSE associate partner).
- Arises from an existing consortium of academic data support professionals working in the domain of social science datasets (Data Information Specialists Committee-UK). We are working together with colleagues engaged in managing open access repositories for e-prints.
- Our project supports academics who wish to openly share datasets and presents a model for depositing ‘orphaned datasets’ that are not being deposited in subject-domain data archives/centres.
- Outputs from the project are intended to help to demystify data as complex objects in repositories, and assist other institutional repository managers in overcoming barriers to incorporating research datasets.
- [http://www.disc-uk.org/datashare.html](http://www.disc-uk.org/datashare.html)
What is a dataset?

- Wikipedia: A data set (or dataset) is a collection of data, usually presented in tabular form. Each column represents a particular variable. Each row corresponds to a given member of the data set in question. It lists values for each of the variables, such as height and weight of an object or values of random numbers. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows.

- DISC-UK: By data, we do not mean a synonym for information. We mean research data, that which is collected, observed, or created, for purposes of analysing to produce original research results. This differs from what is commonly called research outputs, which are the peer reviewed, published papers/articles/books/presentations that are produced as a result of data analysis.

- DISC-UK: Datasets (or data sets) are a group of data files in any format along with the documentation files (such as a codebook, technical report, methodology) which explain their production or use. Generally a dataset is un-usable by a second party unless both parts are included.
Key DC terms for describing datasets

- Type (dataset? collection? others?)
- Format (and hasFormat)
- Coverage (and refinements spatial and temporal)
- Creator, Publisher (how to describe agents?)
- Rights (and refinements accessRights and license)
- Provenance, Source (esp. for derived data)
- Subject (what controlled vocabulary?)
- Relation (isReferencedby, isVersionOf, etc)
Other Relevant Work

- **DRYAD**, UNC-Chapel Hill
  
  http://datadryad.org/about.html

- Application Profile for the eBank UK project and service
  
  http://www.ukoln.ac.uk/projects/ebank-uk/schemas/profile/

- **DDI** versions 2 and 3, http://www.ddialliance.org/


  http://www.disc-uk.org/docs/DDI_and_IRs.pdf

- **GAP**, EDINA and Simon Cox, CSIRO

  http://www.ukoln.ac.uk/repositories/digirep/index/Geospatial_Application_Profile

- **STFC** (formerly CCLRC) Scientific Metadata Model

  http://epubs.cclrc.ac.uk/work-details?w=30324
Motivation for Dryad

- Small science repositories (SSR)
  - Knowledge Network for Biocomplexity (KNB), Marine Metadata Initiative (MMI)
- Evolutionary biology
  - Publication process
    - Supplementary data (*Evolution, American Naturalists*)
      “Author,” “deposition date,” not “subject” “species,” ”geo. locator”
    - Data deposition (Genbank, TreeBase, Morphbank)
- NESCent & SILS/Metadata Research Center
Dryad Repository model
HIVE model

Vocabularies in SKOS format:
- NBII
- LCSH
- TGN
- MeSH

HIVE Vocabulary Server
- SKOS/XML Integration Layer

Transformation Layer

Client Requests

Client
Jörg Holetschek
Botanic Garden & Botanical Museum Berlin-Dahlem
Dept. Biodiversity Informatics & Laboratories
Königin-Luise-Straße 6-8
14195 Berlin

International Conference on DublinCore and Metadata Applications, Berlin, 25.09.08
What’s GBIF/BioCASe?

**GBIF**: Global Biodiversity Information Facility  
**BioCASe**: Biological Collection Access Service  
→ Aim at making the world’s primary biodiversity data freely available on the Internet (for both humans and machines)

**Primary biodiversity data:**  
- Natural history collections  
  (preserved specimens: stuffed, dried, alcoholized; paleontological)  
- Herbaria  
- Living collections (botanical & zoological gardens)  
- Culture collections  
- Observational databases  
  (human observations, Drawings, Photos, Videos, Sounds)

Est. 50-100,000 natural history collections with ~2 billion specimens
Data hierarchy

Currently 147 million records available at GBIF

Institution name, contact information
Address, telephone, eMail, Website

Technical Contact Information
Version Info
Modification Dates

Dataset Title and Description
Geographic and taxonomic coverage
Content Contact Information
Statements (Copyright, terms of use, acknowledgements, citation guidelines, IPR declarations)

Identification (taxon name, classification, identifier, date, references), Geography (continent, country, town, locality)
Gathering date/agent, preparation information
Specimen Details (sex, life stage), multimedia objects
ABCD (Access to Biological Collection Data): XML schema

- <Gathering>
  - <DateTime>
    - <ISODateTime>2008-09-26T00:00:00Z</ISODateTime>
  </DateTime>
- <Agent>
  - <GatheringAgent>
    - <AgentText>Bernd Holub, Maxym M. Boe</AgentText>
  </GatheringAgent>
  - <Method>Collected by hand</Method>
  - <LocalityText>
    - Alpe Maritima, Estriaco (=9km SW of Torino), Valleone del Sabbone (=7km SE of Estriaco)
  </LocalityText>
  - <Country>
    - <Name language="en">Italy</Name>
  </Country>
  - <NamedArea>
    - <AreaName>Parco Naturale delle Alpi Maritime</AreaName>
  </NamedArea>
  - <SiteCoordinateSet>
    - <SiteCoordinate>
      - <CoordinateMethod>Garmin Geo 101 GPS device</CoordinateMethod>
      - <LongitudeDecimal>7.458228</LongitudeDecimal>
      - <LatitudeDecimal>44.175517</LatitudeDecimal>
      - <SpatialDatum>WGS84</SpatialDatum>
      - <AccuracyStatement>20</AccuracyStatement>
      - <CoordinateErrorDistanceInMeters>25</CoordinateErrorDistanceInMeters>
      </SiteCoordinate>
    </SiteCoordinateSet>
  - <Altitude>
    - <MeasurementOrFactAsserted>
      - <LowerValue>1590</LowerValue>
      - <UpperValue>1590</UpperValue>
      </MeasurementOrFactAsserted>
  </Altitude>
- <Metadata>
  - <Description>
    - <RepresentaionLanguage>en</RepresentaionLanguage>
    - <Title>
      - Alpe Maritima Biodiversity Inventory - Mercanton-Alpi Maritima (Torino)</Title>
    - <URI>http://www.ath.s.uw.ac/URI</URI>
    - <Representation>
      - <ImageURI>http://www.ath.s.uw.ac/i mage/logo/log logo edit.png</ImageURI>
    - <Version>
      - Major:0</Major>
      - Minor:0</Minor>
    </Version>
    <RevisionData>
      - <DateCreated>2008-09-26T00:00:00Z</DateCreated>
      - <DateModified>2008-09-26T00:00:00Z</DateModified>
    </RevisionData>
  </Owner>
    - <Organization>
      - <Name>
        - <RepresentationLanguage>en</RepresentationLanguage>
        - <Text>Staatliches Museum für Naturkunde Stuttgart</Text>
        - <Abbreviation>Staatliches Museum für Naturkunde Stuttgart</Abbreviation>
      </Name>
    </Organization>
    - <Person>
      - <Fullname>Christoph Hörer</Fullname>
    </Person>
    - <Roles>
      - <Role>EEM-WF1 Leader</Role>
    </Roles>
    - <Addresses>
      - <Address>Rosental Str. 1, 70191 Stuttgart, Germany</Address>
    </Addresses>
  </Owner>
- <PhoneNumbers>
  - <PhoneNumber>
    - <Number>+49 711 19 36 233</Number>
  </PhoneNumber>
</PhoneNumbers>
Metadata/Data view on GBIF data portal

Actions for Botanic Garden and Botanical Museum Berlin-Dahlem

Explore: Occurrences
Download: Darwin Core records, One-degree cell density overlay for Google Earth, Placemarks for Google Earth (limit 10,000)

Occurrence overview

This map only shows records with coordinates (97,712 records from a total of 286,366 records). It includes records from all datasets shared by this data provider.

Indexed data

Datasets: 22
Occurrences records indexed: 97,712

Note: GBIF, Open Geospatial Consortium services
How To Find Records on GBIF?

- Via **unit data**: taxon name/group
  - (e.g. family „Poaceae“)
  - geospatial (e.g. „Germany“)

- Via **metadata**
  - Dataset title („Fungi“, „Humboldt“)
What is data, what is metadata?

**Example:** Geographic scope

Dataset „VegetWeb: zentrale Datenbank der Arbeitsgemeinschaft Vegetationsdatenbanken; Teil des Netzwerks für Phytodiversität Deutschland (NetPhyD)“ with ~ 250,000 records

No Country Data → cannot be found via geospatial search

Wouldn’t be part of the BioCASe portal for European Botany (http://search.biocase.org/europe)
→ manually tagged with Country = „German“
→ will be included in BioCASe European portal

Is the tag „Germany“ still metadata? Has it become data? Is it both?

**Compilation of metadata is often neglected!**
Metadata often neglected

**Information**

- **Name:** FishBase DiGIR Provider - Philippine Server
- **Website:** www.fishbase.org

**Information**

- **Name:** Desmidiaceae Engels
- **Description:** A new record of reference material of the Desmidiaceae of Germany was compiled, using the databank "Specify". The aim of the project was to summarize and digitalize the data of the dried and otherwise preserved specimens stored in the German herbaria as well as of the living strains, cultured in culture collections. Many of the references are completed by figures of the labels or by scanned micrographs from preparations for the light- or scanning microscopes.
- **Rights:** The use of the data is allowed only for non-profit scientific use and for non-profit nature conservation purpose. The database or part of it may only be used or copied by the written permission from the legal owner. If used for publication, we ask for a copy or an off-print. No part of this data base may be copied or reproduced without written permission from the legal owner.
- **Citation:** Engels, Monika 2003 - (continuously updated) Catalogus novus et amplificatus specimen et viventium algarum Desmidiacearum (New and extended catalogue of herbarium specimen and living material of Desmidiaceae in Germany).

**Information**

- **Name:** Herbarium des Staatlichen Museums für Naturkunde Görlitz (GLM)
- **Description:** The "Herbarium Lusaticum" represents the flora of Upper Lusatia with 47,000 specimens of vascular plants collected over a period of 200 years. About 4,500 specimens of the "Herbarium Lusaticum" are digitised in the collection database.
- **Citation:** Staatliches Museum für Naturkunde Görlitz 1992 - (continuously updated): Vascular Plant Herbarium.
- **How to cite this dataset:** Botanic Garden and Botanical Museum Berlin-Dahlem, Herbarium des Staatlichen Museums für Naturkunde Görlitz (GLM) (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/1105, 2008-08-24)
Assessing Descriptive Substance in Free-Text Collection-Level Metadata

Oksana L. Zavalina, Carole L. Palmer, Amy S. Jackson, Myung-Ja Han

Center for Informatics Research in Science and Scholarship (CIRSS)

Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

8th International Conference on Dublin Core and Metadata Applications
Berlin, September 24, 2008
Digital Collections and Content (DCC) Project

• 2002 – initial IMLS National Leadership Grant; 2005 – grant extension
  • Create an aggregation of cultural heritage digital content
  • How collections and items can best be represented to meet the needs of service providers and diverse user communities.

• 2007 – new IMLS grant
  • Expand the collection/aggregation for targeted scholarly communities based on formal evaluation
  • Develop guidelines for “federation” development
  • Analyze relationships between collection-level metadata and item-level metadata to better preserve context and enhance functionality
DCC Aggregation of Digital Content

- Currently -- over 200 cultural heritage collections
- Adding 140+ collections from ASHO

  - Metadata repository:
    - Harvested metadata aggregated in one location
    - Acts as a portal to the item-level records for digital content in NLG/LSTA collections

  - Collection Registry:
    - Provides access, services, and additional functionality to a database of collection descriptions
    - Collection-level metadata schema adapted in 2003 from a preliminary version of DC CDAP and RSLP (Research Support Libraries Programme, UK)
DCC Collection Metadata Schema

Available at: http://imlsdcc.grainger.uiuc.edu/CDschema_elements.asp
DCC Collection Metadata Schema

• Describes 4 entities:
  – the digital collection
  – the grant project responsible
  – the institution responsible
  – the person(s) responsible for administration of collection

• 30 attributes/elements used for describing the collection:
  – 17 general attributes (title, size, objects represented, language …)
  – 5 topical (GEM subject, [alternative] subject, [free-text] description, geographic coverage, and time period)
  – 4 for relationships with other collections (parent collection, sub-collection, source physical collection, and other associated collection)
  – 4 for relationships with projects, institutions, and administrators (grant project, hosting institution, contributing institution, and administrator).
Full Description of Indian Peoples of the Northern Great Plains

Find out more about: Collection Information
Collections Associated with this Collection
IMLS Grant Projects Responsible for this Collection

Collection Information

Title: Indian Peoples of the Northern Great Plains
URL: http://www.lib.montana.edu/epubs/nadbl/3+

Description: Images of the Indian Peoples of the Northern Great Plains is a searchable online photograph database. The Project strives to broaden access to new constituencies by providing students, researchers, and the general public with direct access to important primary source material on the Plains Indian cultures currently only available by travel to Montana. Images were digitized and drawn from the library collections of three of the Montana State University campuses (Bozeman, Billings, and Havre), the Museum of the Rockies in Bozeman, and Little Big Horn College in Crow Agency, Montana. The digital collection was created in consultation with Native Americans, educators, librarians, and historians. The overall organization of the database is by tribe, including: Crow, Chyowa, Blackfeet, Salish (Flathead), Kutenai, Chippawa-Cree, Gros Ventres (Atsina), and Assiniboin. The collection consists primarily of images, but includes some text to give context. Most of the images are photographs, but there are also stereographs, ledger drawings, and other sketches.

GEM Subjects: Social Studies
Anthropology
Human relations
State history
United States history

Subjects: Native Americans

Geographic Coverage: Mountain Region U.S. (general region)

Time Period: 1350-1399
1000-1020
1330-1449
1870-1944

Objects Represented: Photographs / slides / negatives
Prints and drawings
Treaties

Format: image/gif
image/jpeg
DCC collection-level record example

Language: en

Audience: General public
- Genealogists/historians
- Enthusiasts
- K-12 students
- Undergraduate students
- Staff at peer/partner organizations
- K-12 teachers and administrators
- Scholars/researchers/graduate students

Interaction with Collection: Search

Copyright & IP rights: Digital image files from the database may be used for educational and research purposes only. Commercial publication or reproduction use is prohibited without express written consent from the appropriate collection.

Size: 1,500

Frequency of additions: Irregularly

Metadata schema used: Dublin Core (simple or qualified)

Hosting Institution: Montana State University Libraries

Contributing Institution: Museum of the Rockies

Contributing Institution: Montana State University, Northern, Special Collections and Archives

Contributing Institution: Montana State University, Billings, Library, Special Collection

Contributing Institution: Little Big Horn College

Associated Collections
- There are no associated collections.

IMLS Grant Projects Responsible for Digital Collection

Title of Project: Indian Peoples of the Northern Great Plains

IMLS Grant Type: NLG

IMLS Grant Number: LL-80101
This study aims to:

1. Identify the range of substantive and purposeful information about **collections** available within the DCC Collection Registry

2. Determine patterns of representation

3. Assess the adequacy of the DCC collection-level metadata schema for representing the richness and diversity of collections in the aggregation
Why are we doing this?

• To extend our understanding of the role of collection-level metadata

• To provide an empirical foundation for an ongoing analysis of item-level and collection-level metadata relationships
Content analysis of 202 collection-level records

- Qualitative and quantitative analysis of free-text Description field to identify:
  - types of information provided about a digital collection (collection properties)

- Grounded approach (properties emerged from coding; intercoder reliability of 80.4% agreement in assigning the codes to specific cases)

- 14 collection properties found in 5% or more collection records

- degree of agreement/overlap with information provided in other free-text and controlled-vocabulary collection metadata fields
Additional analysis

– 4 collection-level metadata fields intended for subject indexing:
  • GEM Subjects
  • [alternative] Subjects
  • Geographic Coverage
  • Time Period

– The field describing types of objects in digital collections
  • Objects Represented
Analyzed metadata fields at a glance

**Description:** Images of the Indian Peoples of the Northern Great Plains is a searchable online photograph database. The Project strives to broaden access to new constituencies by providing students, researchers, and the general public with direct access to important primary source material on the Plains Indian cultures currently only available by travel to Montana. Images were digitized and drawn from the library collections of three of the Montana State University campuses (Bozeman, Billings, and Havre), the Museum of the Rockies in Bozeman, and Little Big Horn College in Crow Agency, Montana. The digital collection was created in consultation with Native Americans, educators, librarians, and historians. The overall organization of the database is by tribe, including: Crow, Cheyenne, Blackfeet, Salish (Flathead), Kutenai, Chippewa-Cree, Gros Ventres (Atilna), and Assiniboine. The collection consists primarily of images, but includes some text to give context. Most of the images are photographs, but there are also stereographs, ledger drawings, and other sketches.

**GEM Subjects:** Social Studies
- Anthropology
- Human relations
- State history
- United States history

**Subjects:** Native Americans

**Geographic Coverage:** Mountain Region U.S. (general region)

**Time Period:**
- 1850-1899
- 1900-1929
- 1930-1949
- 1870-1954

**Objects Represented:**
- Photographs / slides / negatives
- Prints and drawings
- Treaties
Collection properties found only in the \textit{Description} field

<table>
<thead>
<tr>
<th>Collection Property</th>
<th>Number of collections</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GROUP 1 (&quot;special claims&quot;)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Importance</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>Uniqueness</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>Comprehensiveness</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td><strong>GROUP 2</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item Creator</td>
<td>78</td>
<td>39</td>
</tr>
<tr>
<td>Provenance</td>
<td>24</td>
<td>12</td>
</tr>
<tr>
<td><strong>GROUP 3</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subjects not represented in formal metadata elements</td>
<td>132</td>
<td>67</td>
</tr>
<tr>
<td>Objects not represented in formal metadata elements</td>
<td>37</td>
<td>19</td>
</tr>
</tbody>
</table>
Features of interest to scholarly audiences

Not represented elsewhere in collection records

• **Group 1.** “Special claims” about a collection:
  
  – *Importance, Uniqueness, and Comprehensiveness*

• Add vital qualitative, contextual information about:
  – intentions of collectors
  – role the collection plays in the larger universe of related content

• Correspond to *Strength* collection metadata element:
  – present in RSLP collection description schema
  – discussed in DC CDAP community several years ago.
Some examples of “special claims”

- “Collection of the most important and influential 19th and early 20th century American cookbooks”

- “Materials are significant in their place within the fabric of American history and culture”

- “Unique historical treasures from ... archives, libraries, museums, and other repositories”

- “Rare and unique library and archival resources on race relations”

- “A comprehensive and integrated collection of sources and resources on the history and topography”

- “One of the most ambitious and comprehensive effort to date to deliver educational content on the Civil Rights Movement”
### Collection properties found only in the Description field

<table>
<thead>
<tr>
<th>Collection Property</th>
<th>Number of collections</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GROUP 1 (&quot;special claims&quot;)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Importance</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>Uniqueness</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>Comprehensiveness</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td><strong>GROUP 2</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item Creator</td>
<td>78</td>
<td>39</td>
</tr>
<tr>
<td>Provenance</td>
<td>24</td>
<td>12</td>
</tr>
<tr>
<td><strong>GROUP 3</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subjects not represented in formal metadata elements</td>
<td>132</td>
<td>67</td>
</tr>
<tr>
<td>Objects not represented in formal metadata elements</td>
<td>37</td>
<td>19</td>
</tr>
</tbody>
</table>
Features of interest to scholarly audiences

Not represented elsewhere in collection records

• **Group 2.** Important properties for which no specific elements in DCC collection metadata exist

  – *Provenance*
    • Covered by *Custodial History* collection metadata element in DC CDAP

  – *Item Creator*
    • Not available in DC CDAP or RSLP collection metadata schemas

    • DC CDAP *Collector* element is designed to cover creator of the collection
**Provenance and Item Creator examples**

**Provenance**

- “Acquisition of these hitherto unknown manuscripts was spearheaded by Edgar J. Goodspeed in the first half of the twentieth century”
- “A 1988 bequest of more than 850 landscape prints and drawings from the collection of Los Angeles architect Rudolf L. Baumfeld significantly enhanced this wide-ranging and well-studied thematic area”

**Item Creator**

- “Images are noted on their mounts as being from Watkins's "New Series".... Watkins was active between 1854 and the late 1890s.”
<table>
<thead>
<tr>
<th>Collection Property</th>
<th>Number of collections</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROUP 1 (&quot;special claims&quot;)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Importance</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>Uniqueness</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>Comprehensiveness</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>GROUP 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item Creator</td>
<td>78</td>
<td>39</td>
</tr>
<tr>
<td>Provenance</td>
<td>24</td>
<td>12</td>
</tr>
<tr>
<td>GROUP 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subjects not represented in formal metadata elements</td>
<td>132</td>
<td>67</td>
</tr>
<tr>
<td>Objects not represented in formal metadata elements</td>
<td>37</td>
<td>19</td>
</tr>
</tbody>
</table>
More features of interest to scholars

- **Group 3. Properties for which formal elements do exist but Description field provides extensive additional coverage**
  
  - *Subjects* and *Object types*

    - Two most widely represented properties (91% and 75%)
    - More accurate in coverage than other metadata fields (67% and 19%)
    - More detail than other fields specified for those purposes (example on the next slide).
Subjects property example

Description:
approximately 150 cubic feet of administrative, survey and fieldwork files and tens of thousands of audio and video recordings dating from the 1930s through 2001. The collection consists of 88 record series documenting performances by, interviews with, and fieldwork surveys of folk musicians, craftspersons, storytellers, folklife interpreters, and cultural tradition-bearers in such areas as children’s lore, foodways, religious traditions, Native American culture, maritime traditions, ethnic folk culture, material culture, and occupational lore.

GEM Subjects:
Arts:
- Architecture
- Music
- Popular culture
- Theater arts
- Visual arts

Educational Technology

Religion

Social Studies:
- State history
- United States history

Geographic Coverage:
- United States (nation)
- Southern U.S. (general region)
- Florida (state)

Time Period:
- 1950-1969
- 1970-1999
- 1930-1949
- 2000 to present
Subjects in Description field

- Content varies:
  - explicit subject coverage statements:
    - “cover a broad range of topics, including ranching, mining, land grants, crime on the border, and governmental issues.”
  - subject keywords scattered throughout the text:
    - “During World War II, as a member of the U. S. Army, 252nd Field Artillery Battalion, he captured over 700 images of life as a soldier and unique snapshots of events of the war”.

- Free-text Description field often adds essential subject information
  - more accurate and specific coverage than fields intended for subject indexing
Objects property example

Description: A unique collection of ephemera, published materials, and artifacts from U.S. national political campaigns (1800-1976). The collection consists of published material, ephemera, and artifacts dating to between 1800 and 1976, including ballots and slates of candidates; promotional broadsides, handbills, and posters; political cartoons (primarily from Harper's Weekly, Frank Leslie's Illustrated Newspaper, and Puck); lithographs and prints (primarily by Kellogg, N. Currier, and Currier & Ives); pamphlets, leaflets, and brochures; songbooks and sheet music; badges, pins, ferrotypes and celluloid buttons; campaign ribbons; parade equipment such as lanterns, torches, banners, and walking sticks; bandanas and other textiles; and souvenirs of all kinds including plates, cups, vases, trays, bottles, sewing boxes, and games.

Objects Represented: Books and pamphlets
Newspapers
Posters and broadsides
Prints and drawings
Physical artifacts
Caricatures
Political cartoons
Cartoons (Commentary)
More complementary contextual information:

- Collection development criteria and title (52% each)
- Collection size (27%)
- **Audiences** (17%)
- Navigation and functionality (16%)
- Participating/contributing institutions (15%)
- Funding sources (5%)
- etc.
Audience: more specific in Description field

**Description:**

Museum of Photography faces the challenge of providing ready, useful and intellectual access to a valuable body of cultural and educational resources of interest to the general public and scholars alike. Consisting of 250,000 stereoscopic glass-plate and film negatives and 100,000 vintage prints,

Collection is the archive of the Keystone View Company of Meadville, PA (active from 1892-1963). As a collection, it is the world's largest body of original stereoscopic negatives and prints providing an encyclopedic view of global cultural history. Formed over the period of the United States' emergence as a world power, not only chronicles an age, it also represents in pictures a dominant point of view about the world during the nineteenth and twentieth centuries. It is an important tool for among others, anthropologists, art historians, cultural studies scholars, historians, political scientists and sociologists. The Keystone-Mast Collection Guide 2003 provides online access to approximately twenty percent of the total stereographic collection. To date, it represents content from the following geopolitical subject areas: entries from North America, from Central America, from West Indies (Caribbean Islands), from South America, from Oceania, from Asia, from Africa, and from the Middle East. When finished, the collection guide will consist of well over 100,000 online stereoviews complete with metadata.

**Audience:**

- General public
- K-12 students
- Undergraduate Students
- K-12 teachers and administrators
- Scholars/Researchers/Graduate Students
Conclusions

• Free-text metadata is as important for collection-level access as controlled-vocabulary metadata
  – one complements the other

• DCC collection metadata schema needs to be:
  – Aligned with current version of DC CDAP:
    • *Custodial History* field will accommodate *Provenance* property currently found only in the *Description* field
  – Updated with newly defined fields for:
    • Creators of items in a collection (*Item Creator?*)
    • Special claims about collections (*Strengths?*)
Conclusions

• Varied use of free-text *Description* field:
  – includes information on institutions, physical and digital collections
  – difficult to automate extraction to populate or enhance other elements

• BUT *Description* field could lend itself to mining:
  – for production of controlled vocabularies customized for use in the DCC and similar aggregations
  – experiment first with improving our existing vocabularies for:
    • *Objects Represented* and *Audience*
    • Possibly subject areas with strong concentrations of content (e.g., Midwest history, American South History, Native Americans history, etc.).
Further research

- Comparative analysis of collection-level records from sources other than the DCC aggregation

- Reproducing user search queries collected through transaction logs:
  - Where in collection-level metadata the matches to user search terms occur?
    
    - What proportion of records retrieved by a keyword search has a keyword only in a free-text Description field and thus would not be retrieved if there were no free-text Description field?

    - What proportion of records retrieved by a keyword search has a keyword only in formal metadata element(s) and thus would not be retrieved if there were no formal metadata element(s) in collection metadata schema?

- User study
More information on DCC website

Digital Collections and Content

Working toward interoperable digital content.

What's Here?

Digital Collections and Content contains descriptions of digital resources developed by IMLS grantees. Examples of what you will find here include info about your watershed from INFOmine at UC Riverside, paintings made during Japanese internment from MOAC, drawings of period dress from Broward County’s Education by Design, Lewis & Clark’s journals from Wisconsin’s American Journeys.

Project News

- IMLS DCC receives continued funding for 2007-2010.
- Currently working on a new search interface.

About the Project

The Digital Collections and Content (DCC) project is investigating and implementing a systematic approach to developing useful, meaningful, and usable digital collections. This collaboration with IMLS and IMLS-funded projects supports IMLS’ mission to create a nation of learners and sustain cultural heritage. Learn more about the project.
Acknowledgements

• This research has been funded by IMLS NLG Research and Demonstration grant LG-06-07-0020-07
  http://imlsdcc.grainger.uiuc.edu/

• Special thanks to:
  • Timothy W. Cole – Principal Investigator
  • Sarah Shreeves – former Project Coordinator
  • Metadata Roundtable members
Questions and comments are always welcome

Oksana L. Zavalina  zavalina@illinois.edu
Carole L. Palmer  clpalmer@illinois.edu
Amy S. Jackson  amyjacks@illinois.edu
Myung-Ja Han  mhan3@illinois.edu
The End

Thank you!