Institute of Museum and Library Services

Digital Collections and Content

Working toward interoperable digital content.

Assessing Descriptive Substance in Free-Text Collection-Level Metadata

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

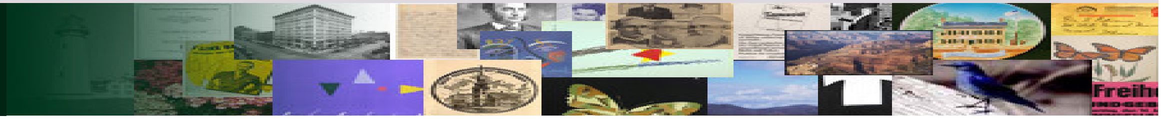
Center for Informatics Research in Science and Scholarship (CIRSS)

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

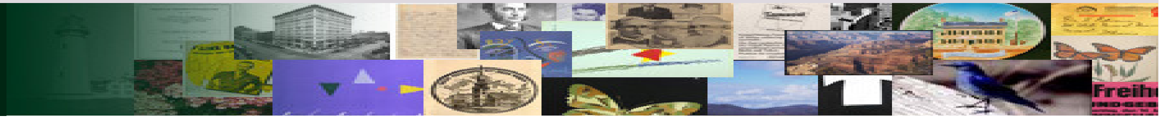
Digital Collections and Content (DCC) Project

- 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



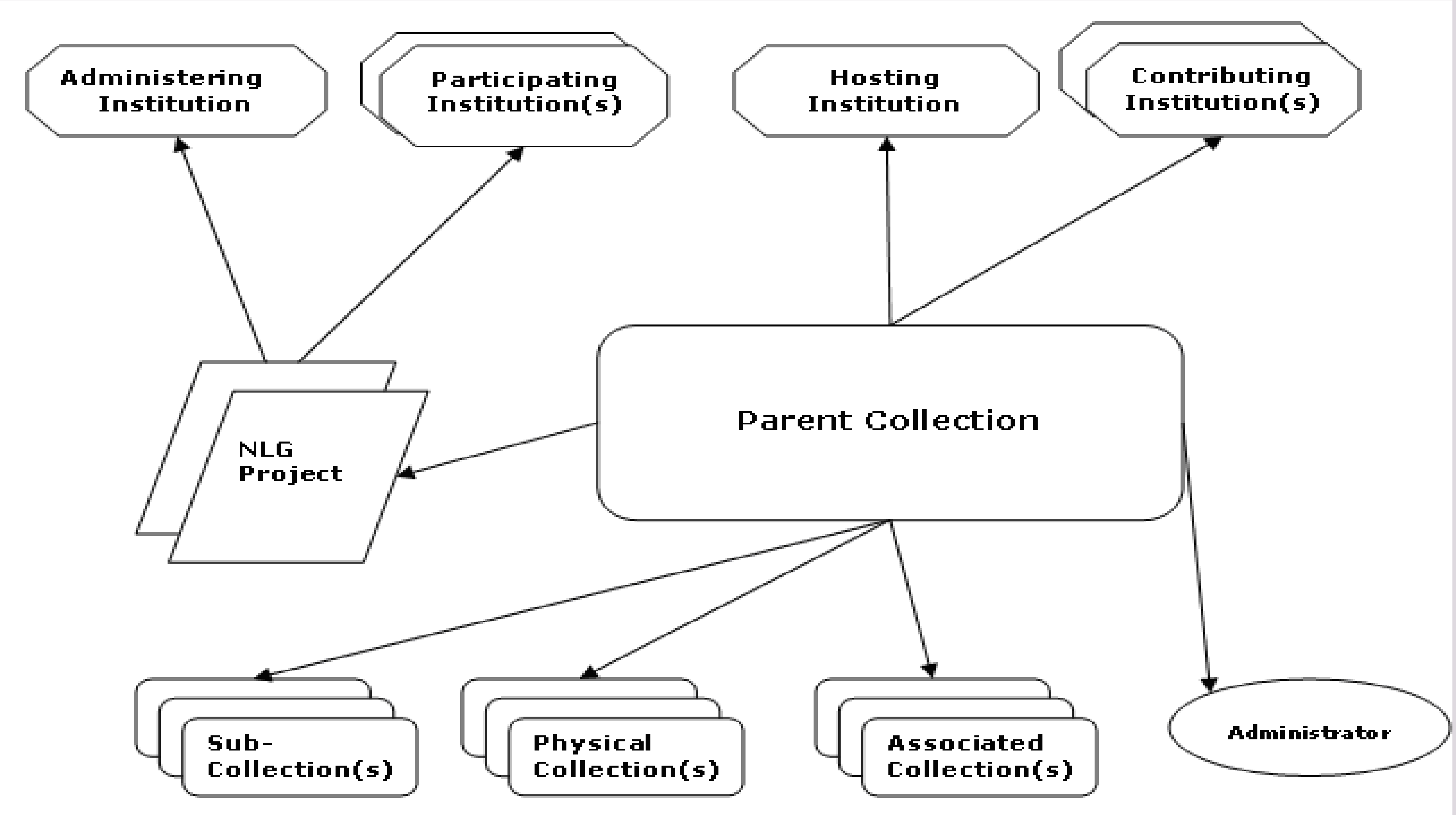
DCC Aggregation of Digital Content

- 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**
    - **<u>Collection-level metadata schema</u> 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

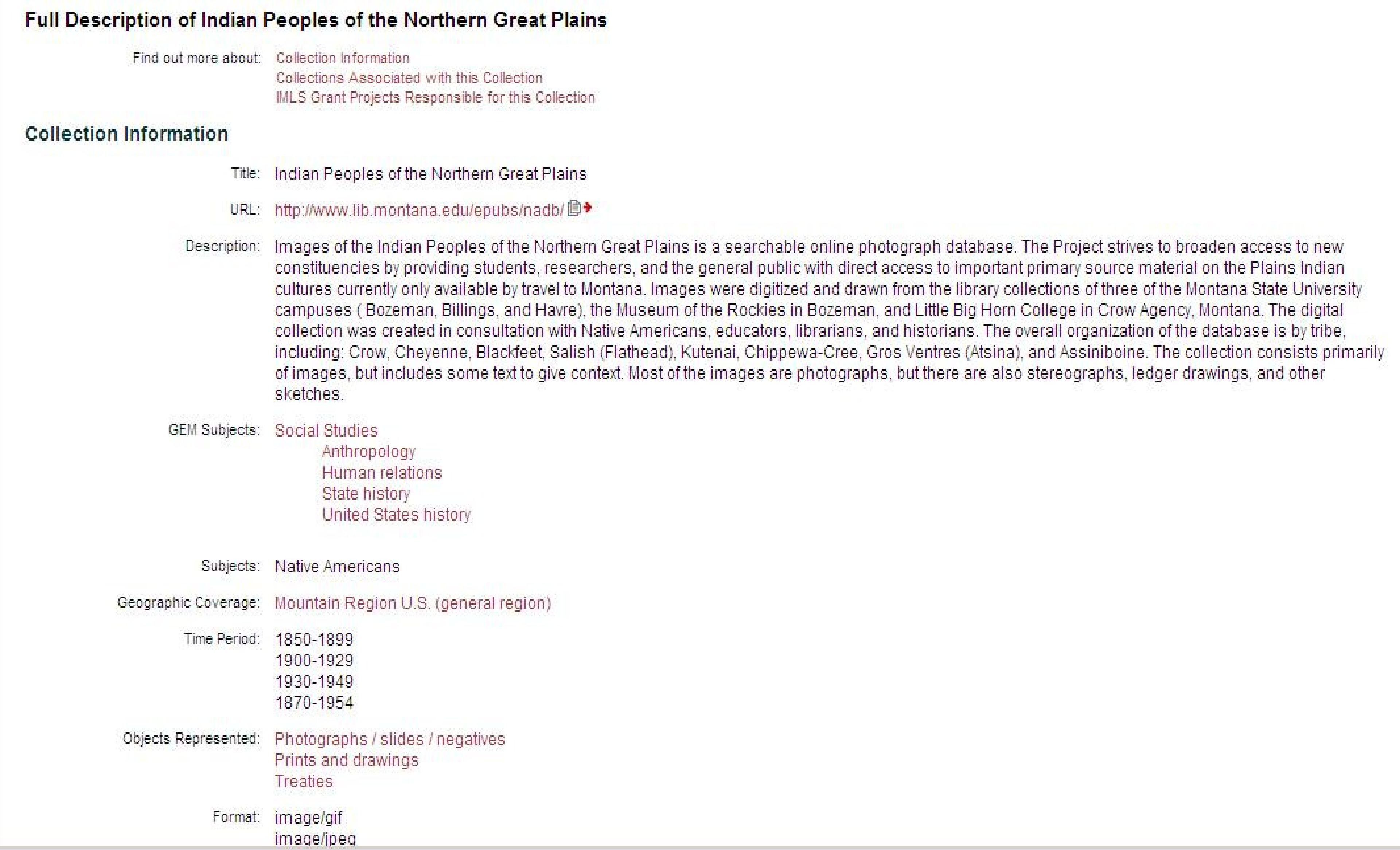


DCC Collection Metadata Schema

- 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).

DCC collection-level record example



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/nadb/

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 (Atsina), 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

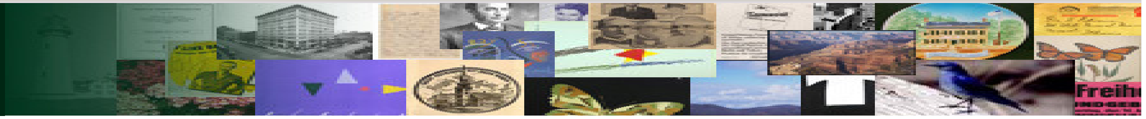
Subjects: Native Americans

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

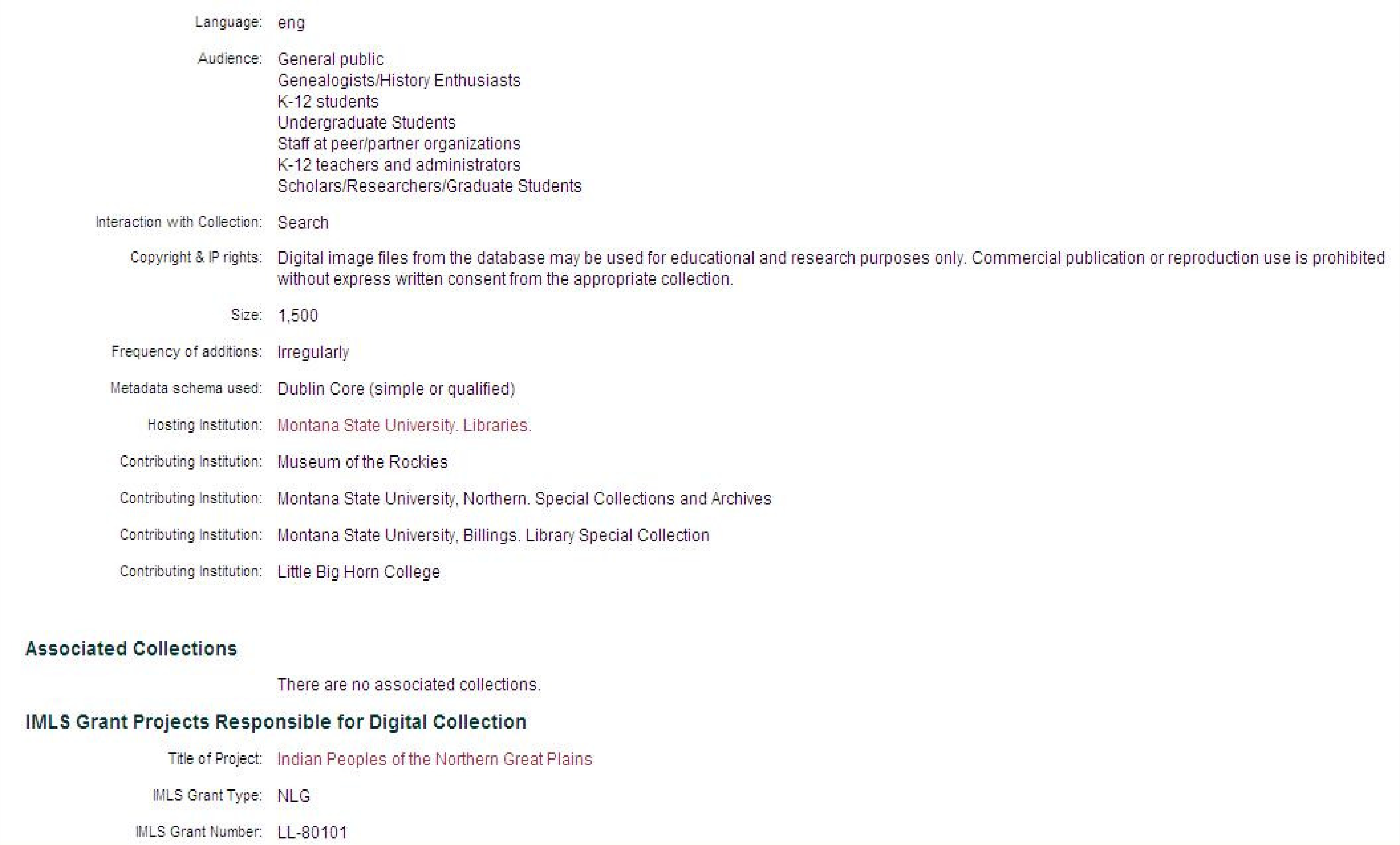
Time Period: 1850-1899
1900-1929
1930-1949
1870-1954

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

Format: image/gif
image/jpeg



DCC collection-level record example



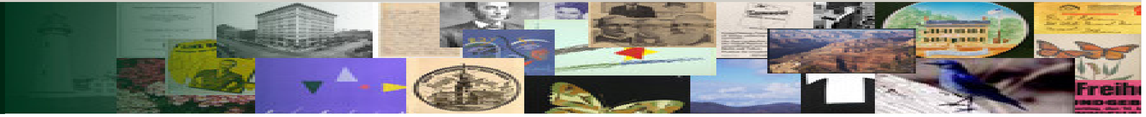
Language:	eng
Audience:	General public Genealogists/History Enthusiasts K-12 students Undergraduate Students Staff at peer/partner organizations K-12 teachers and administrators Scholars/Researchers/Graduate Students
Interaction with Collection:	Search
Copyright & IP rights:	Digital image files from 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

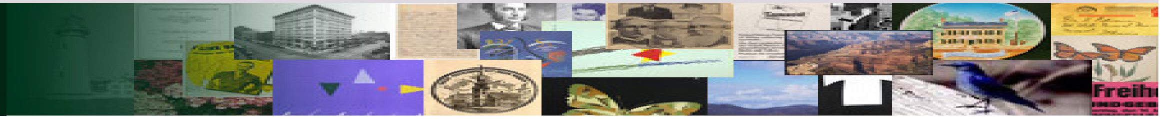


This study aims to:

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

2. Determine patterns of representation

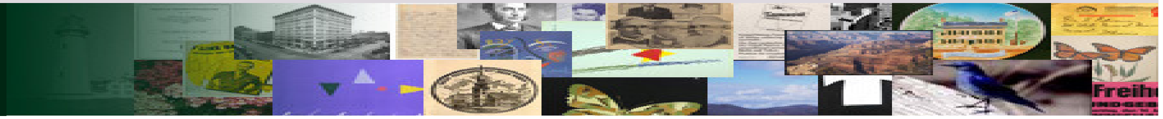
3. Assess the adequacy of the DCC collection-level metadata schema for representing the richness and diversity of collections in the aggregation



Why are we doing this?

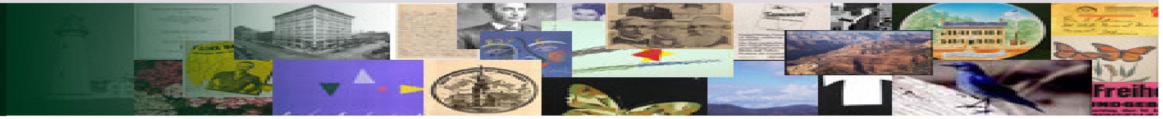
- To extend our understanding of the role of collection-level metadata

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



Content analysis of 202 collection-level records

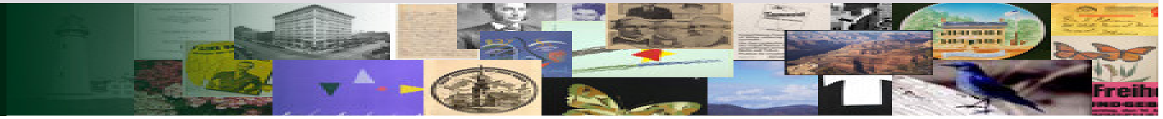
- Qualitative and quantitative analysis of free-text *Description* field to identify:
  - types of information provided about a digital collection (***collection properties***)
    - Grounded approach (properties emerged from coding; intercoder reliability of 80.4% agreement in assigning the codes to 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



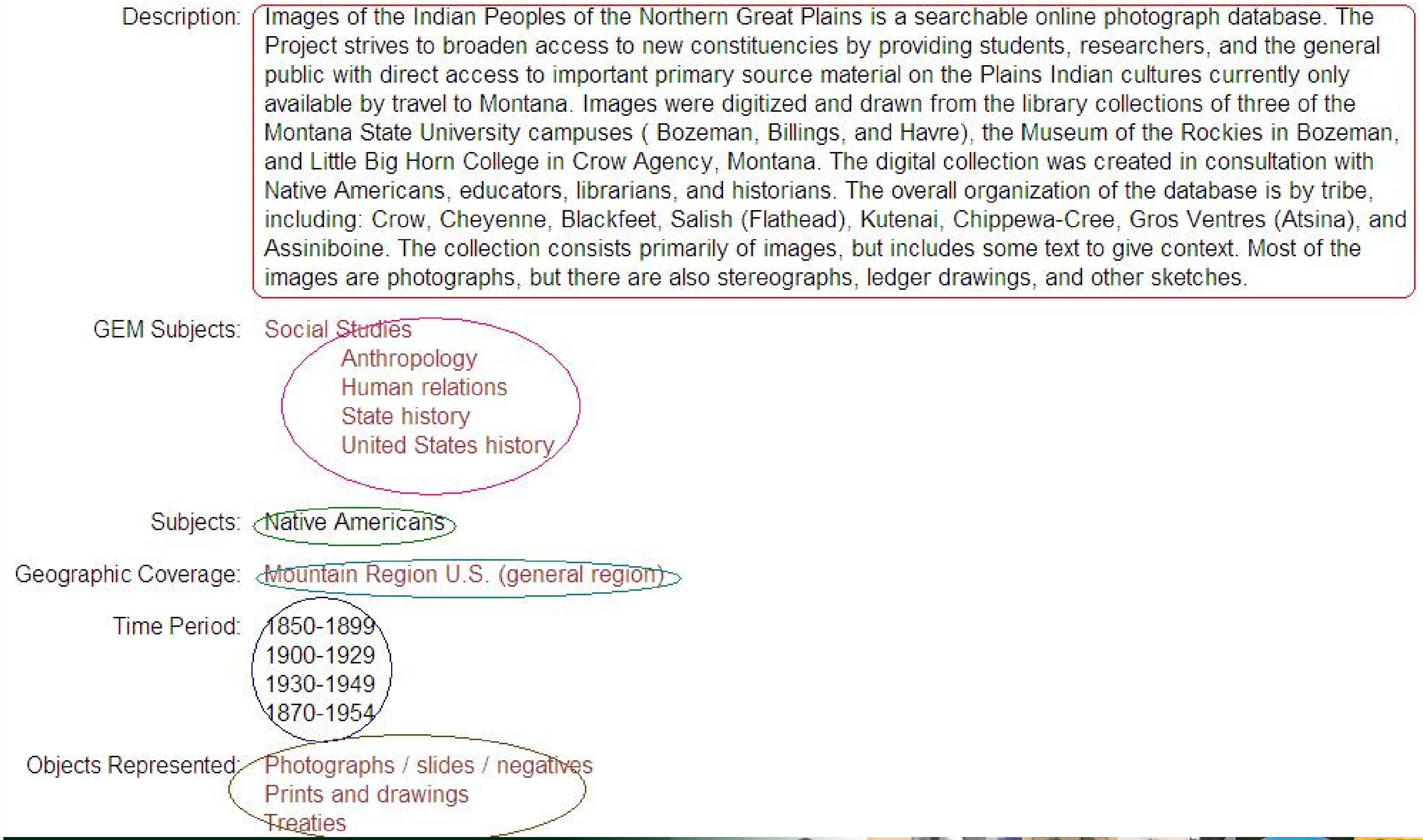
Additional analysis

- 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 (Atsina), 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.

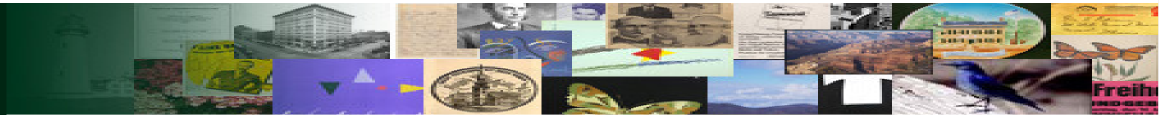
GEM Subjects: Social Studies
- Anthropology
- Human relations
- State history
- United States history

Subjects: Native Americans

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

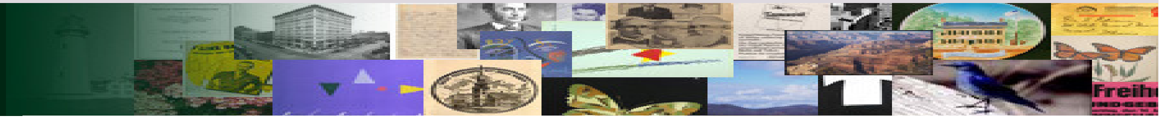
Time Period: 1850-1899
1900-1929
1930-1949
1870-1954

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



Collection properties found only in the *Description* field

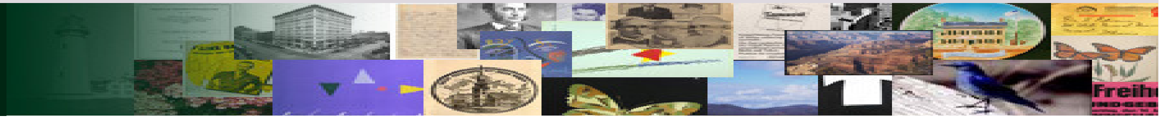
Collection Property	***Number of collections***	***%***
GROUP 1 (“special claims”)		
Importance	20	10
Uniqueness	17	9
Comprehensiveness	6	3
GROUP 2		
Item Creator	78	39
Provenance	24	12
GROUP 3		
Subjects not represented in formal metadata elements	132	67
Objects not represented in formal metadata elements	37	19



Features of interest to scholarly audiences

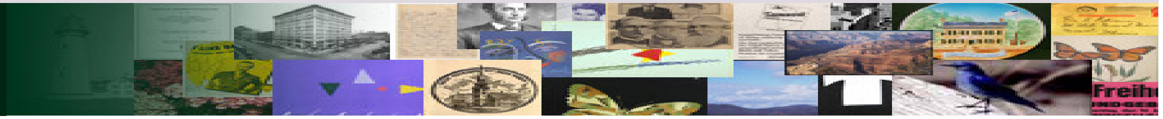
Not represented elsewhere in collection records

- **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.



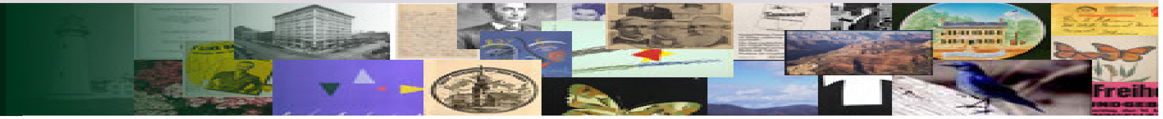
Some examples of “special claims”

- “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

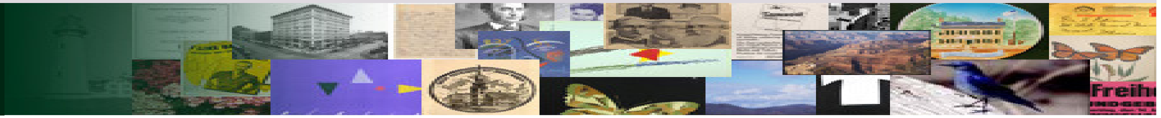
Collection Property	*Number of collections*	*%*
GROUP 1 (“special claims”)		
Importance	20	10
Uniqueness	17	9
Comprehensiveness	6	3
GROUP 2		
Item Creator	78	39
Provenance	24	12
GROUP 3		
Subjects not represented in formal metadata elements	132	67
Objects not represented in formal metadata elements	37	19



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

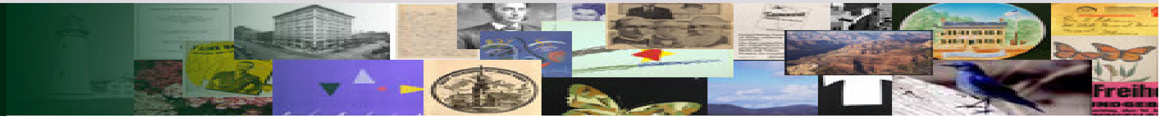
- “Acquisition of these hitherto unknown manuscripts was spearheaded by Edgar J. Goodspeed in the first half of the twentieth century”

- “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

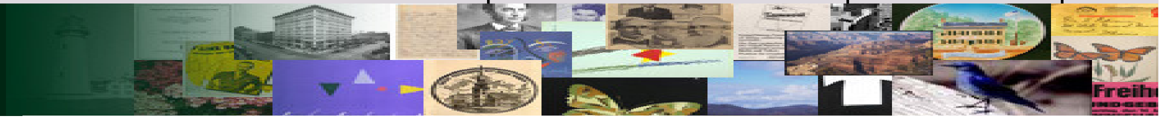
- “The Museum Extension Projects of Pennsylvania, New Jersey, Connecticut, Illinois, and Kansas crafted most of the items currently in the collection”

- “Images are noted on their mounts as being from Watkins's "New Series".... Watkins was active between 1854 and the late 1890s.”



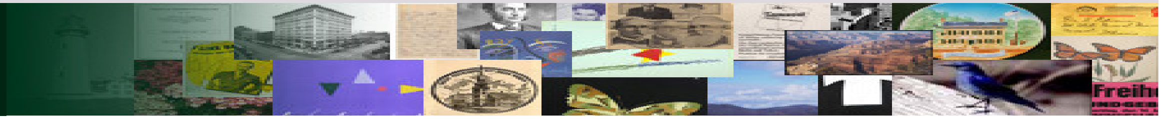
Collection properties found only in the *Description* field

Collection Property	*Number of collections*	*%*
GROUP 1 (“special claims”)		
Importance	20	10
Uniqueness	17	9
Comprehensiveness	6	3
GROUP 2		
Item Creator	78	39
Provenance	24	12
GROUP 3		
Subjects not represented in formal metadata elements	132	67
Objects not represented in formal metadata	37	19

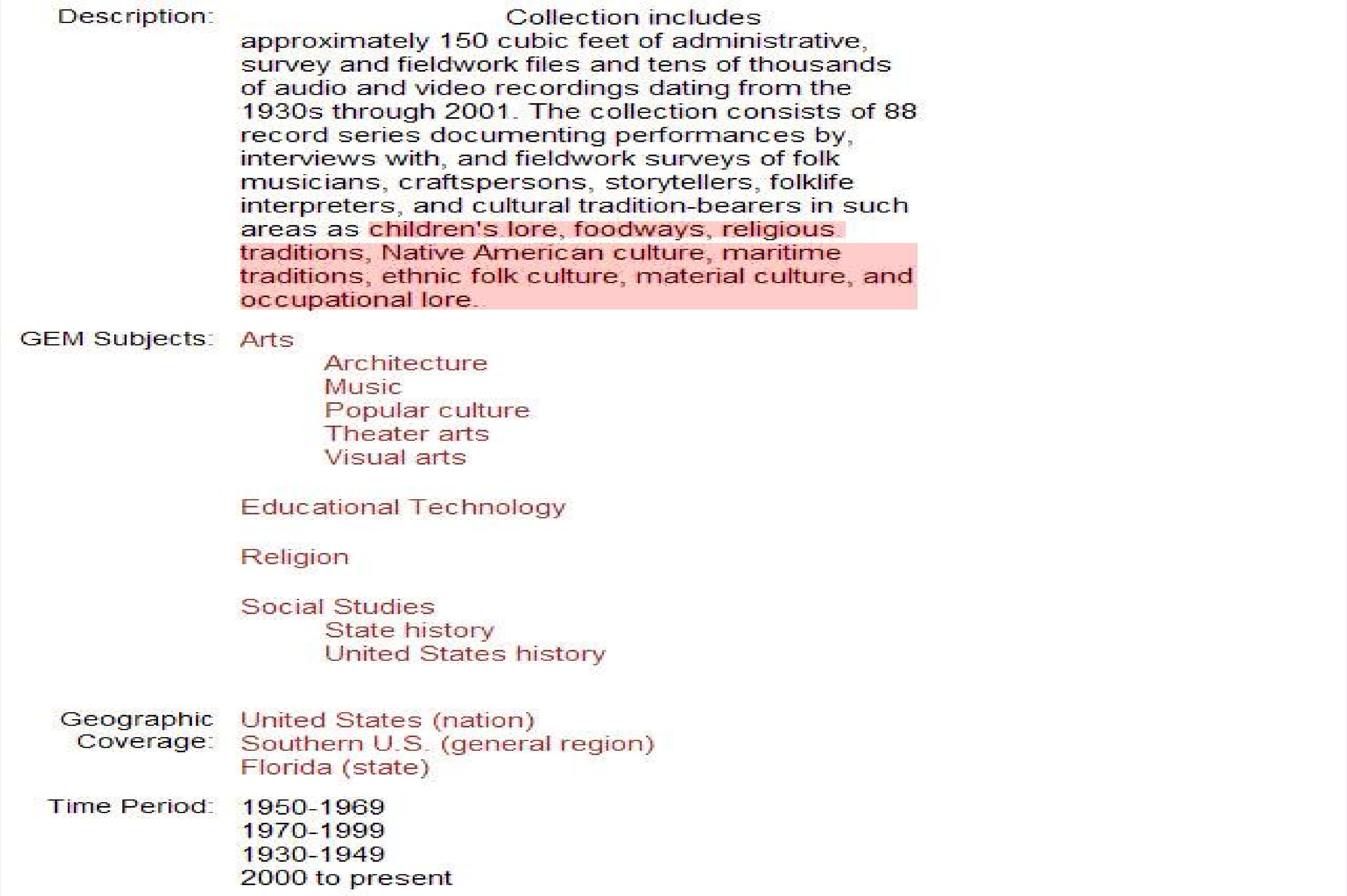


More features of interest to scholars

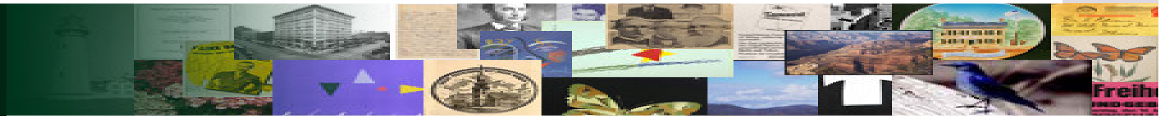
- **Group 3. Properties for which formal elements do exist but Description field provides extensive additional coverage**
  - ***Subjects*** and ***Object types***
    - Two most widely represented properties (91% and 75%)
    - More accurate in coverage than other metadata fields (67% and 19%)
    - More detail than other fields specified for those purposes (example on the next slide).



Subjects property example

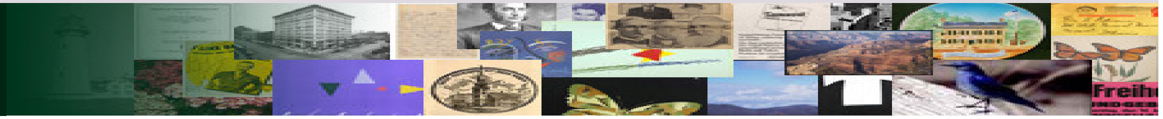


Description:	Collection includes approximately 150 cubic feet of administrative, survey and fieldwork files and tens of thousands of audio and video recordings dating from the 1930s through 2001. The collection consists of 88 record series documenting performances by, interviews with, and fieldwork surveys 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

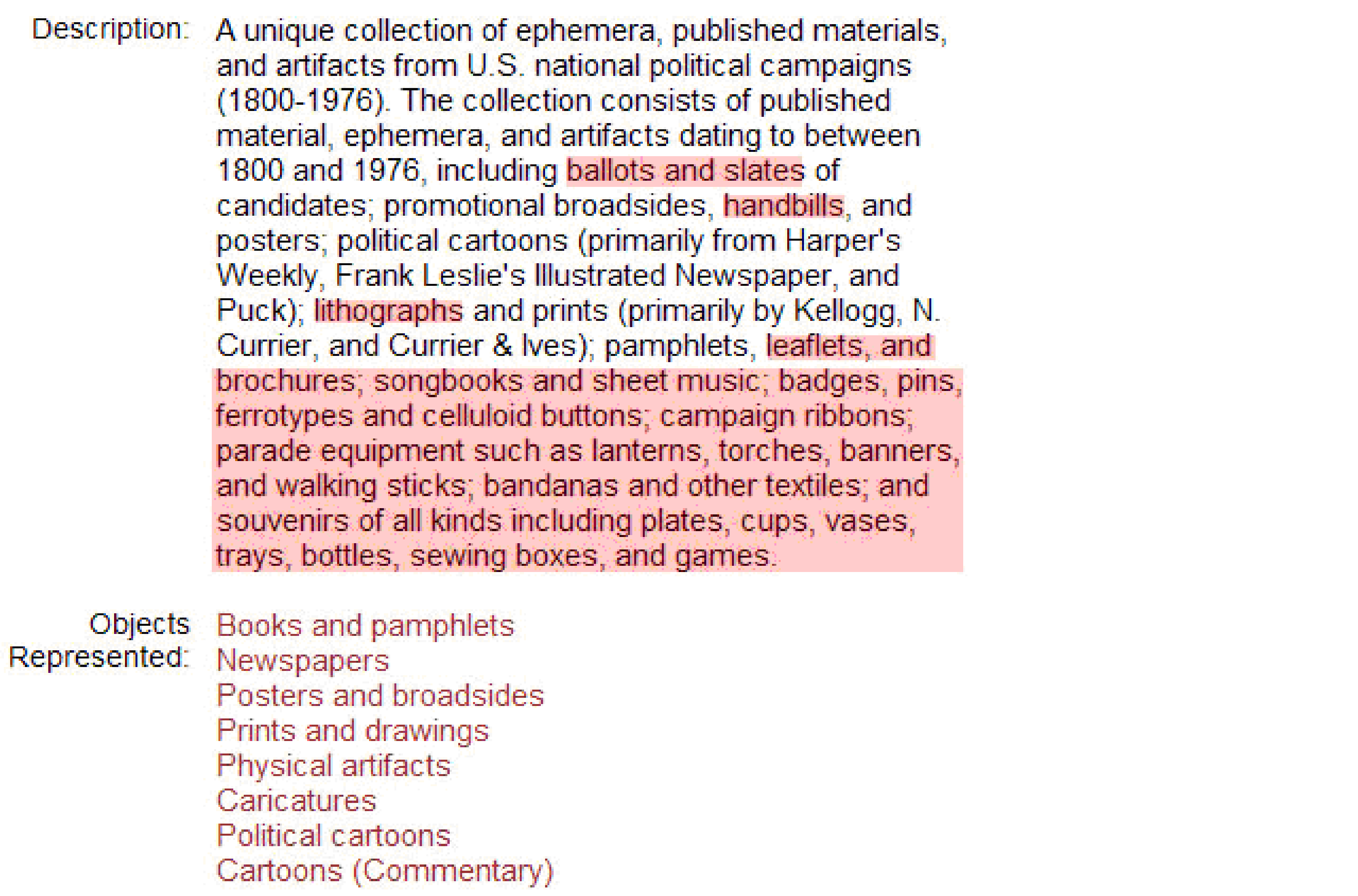


Subjects in Description field

- 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

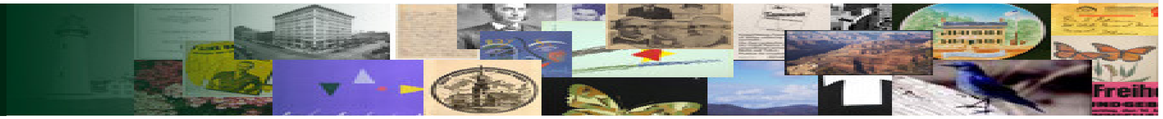


Objects property example



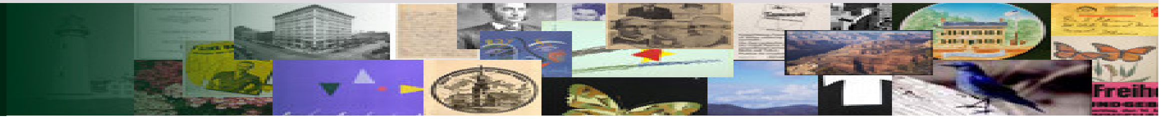
Description: A unique collection of ephemera, published materials, and artifacts from U.S. national political campaigns (1800-1976). The collection consists of published material, ephemera, and artifacts dating to between 1800 and 1976, including ballots and slates of candidates; promotional broadsides, 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)

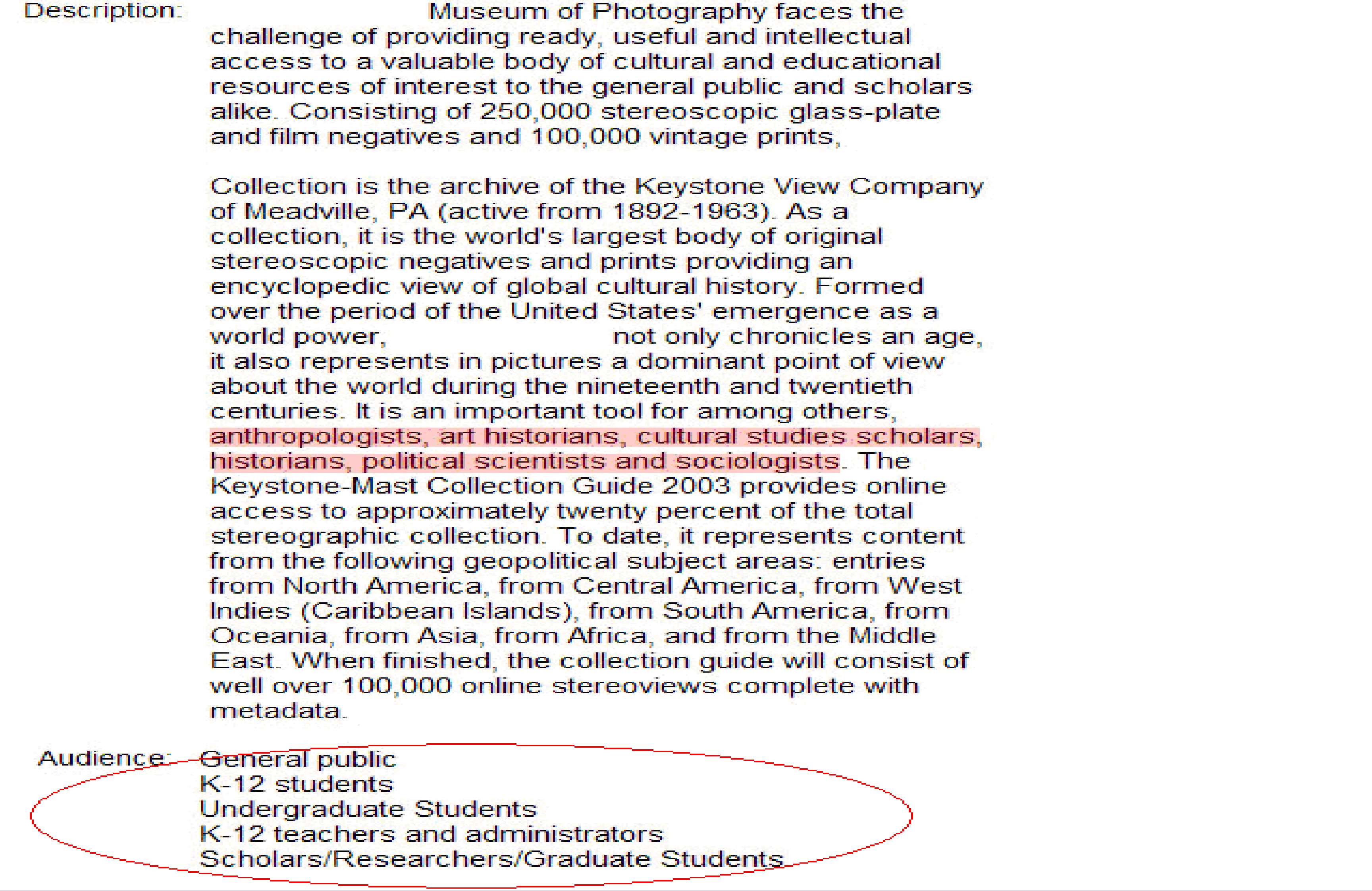


More complementary contextual information:

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



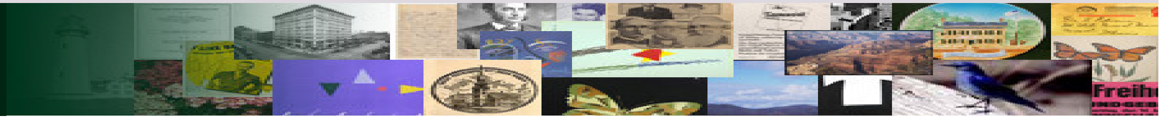
Audience: more specific in *Description* field



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.

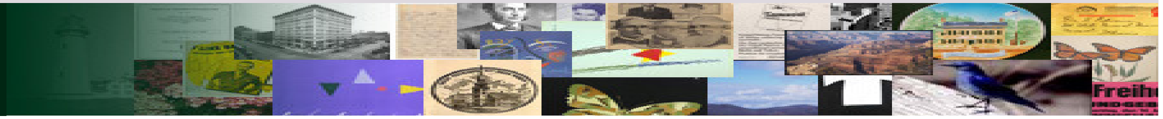
Audience: General public
K-12 students
Undergraduate Students
K-12 teachers and administrators
Scholars/Researchers/Graduate Students



Conclusions

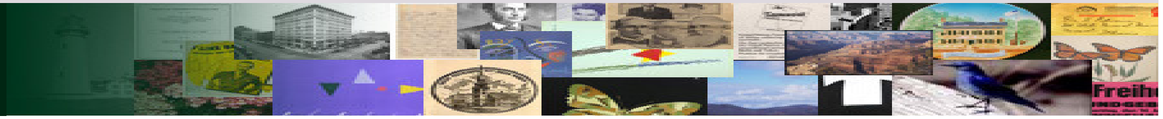
- Free-text metadata is as important for collection-level access as controlled-vocabulary metadata
  - one complements the other

- **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

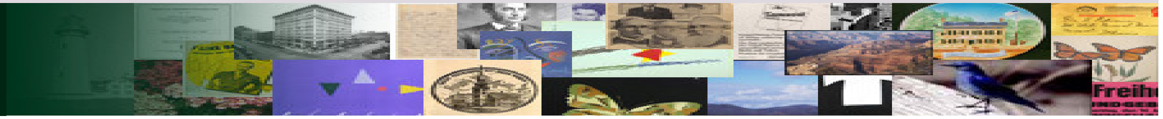
- Reproducing user search queries collected through transaction logs:

  - Where in collection-level metadata the matches to user search terms occur?

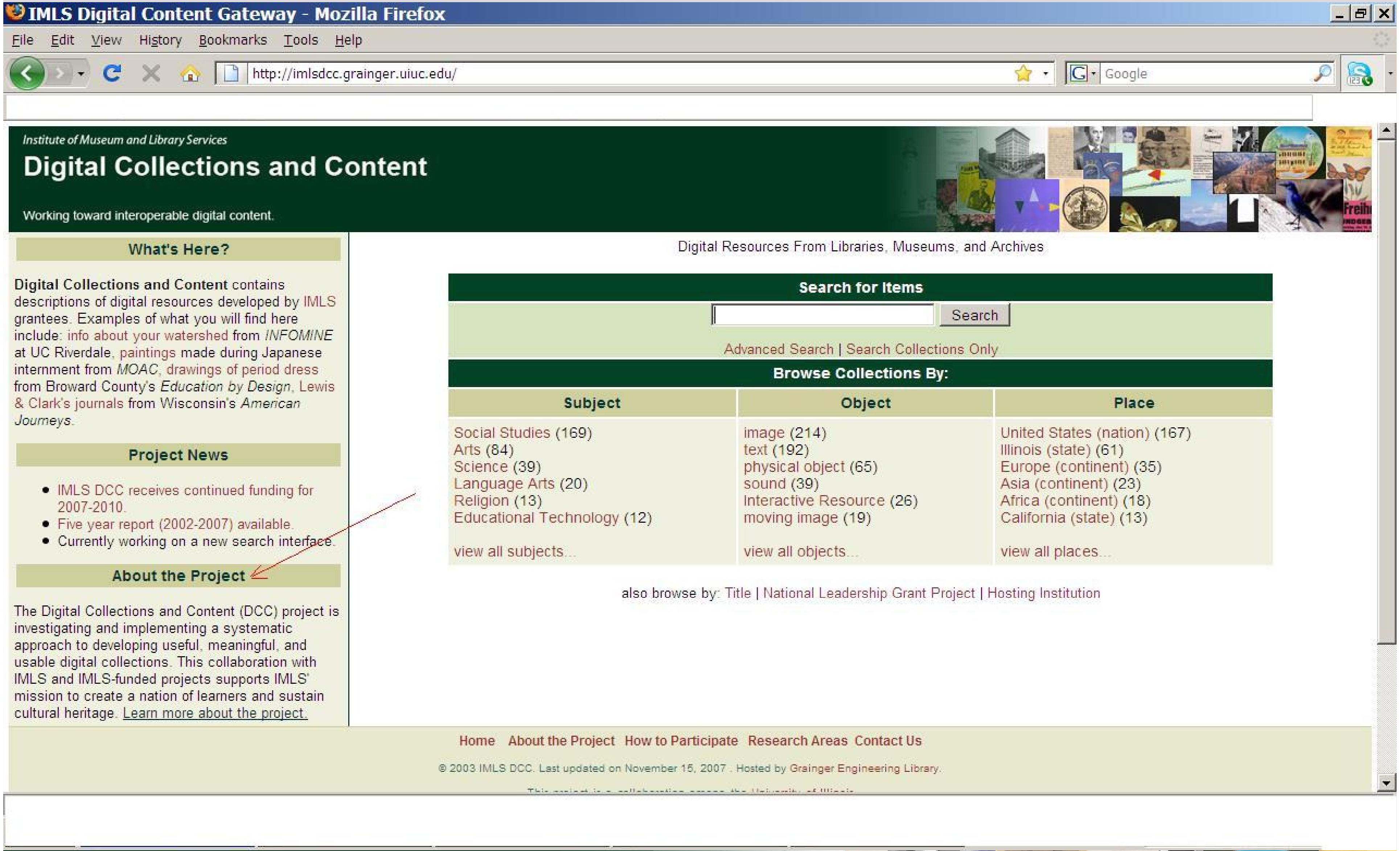
    - What proportion of records retrieved by a keyword search has a keyword only in a free-text *Description* field and thus would not be retrieved if there were no free-text *Description* field?

    - 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


IMLS Digital Content Gateway - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://imlsdcc.grainger.uiuc.edu/
Google

Institute of Museum and Library Services

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 Riverdale, 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.
- Five year report (2002-2007) available.
- 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.

Digital Resources From Libraries, Museums, and Archives

Search for Items

Search

Advanced Search | Search Collections Only

Browse Collections By:

Subject	Object	Place
Social Studies (169)	image (214)	United States (nation) (167)
Arts (84)	text (192)	Illinois (state) (61)
Science (39)	physical object (65)	Europe (continent) (35)
Language Arts (20)	sound (39)	Asia (continent) (23)
Religion (13)	Interactive Resource (26)	Africa (continent) (18)
Educational Technology (12)	moving image (19)	California (state) (13)
view all subjects...	view all objects...	view all places...

also browse by: Title | National Leadership Grant Project | Hosting Institution

Home About the Project How to Participate Research Areas Contact Us

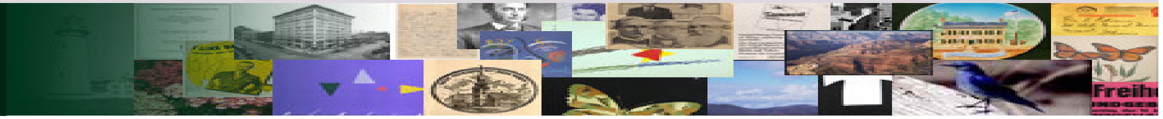
© 2003 IMLS DCC. Last updated on November 15, 2007 . Hosted by Grainger Engineering Library.

This project is a collaboration among the University of Illinois

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



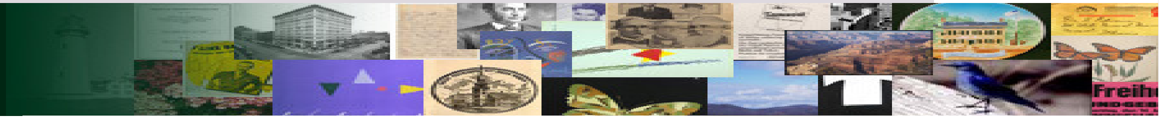
Questions and comments are always welcome

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!