Experiences Using the Open Archives Initiative (OAI)

Martin Halbert - Chris Powell - David Seaman - Stephen Schwartz - Joanne Kaczmarek

Association for Computing and the Humanities
Atlanta, GA June 1, 2003

PANELISTS

- Martin Halbert – Emory University
- Chris Powell – University of Michigan
- Stephen Schwartz – UCLA
- David Seaman – Digital Library Federation
- Joanne Kaczmarek – University of Illinois

BACKGROUND

- “OAI develops and promotes interoperability solutions aimed to facilitate efficient dissemination of content.”
- 1999 Santa Fe Convention – physics e-prints

OAI BACKGROUND

Herbert van de Sompel (LANL) and Carl Lagoze (Cornell)

Two years of funding from the Digital Library Federation (DLF) and the Coalition for Networked Information (CNI)

Mainstay of such services as the NSF’s National Science Digital Library (NSDL)

Much non-US interest and activity

OAI BACKGROUND

- Growing interest from funding agencies – NSF; IMLS; Mellon

- Mellon Foundation funds seven OAI-PMH projects
  - University of Michigan
  - University of Illinois
  - University of Virginia
  - Emory University
  - Solinet
  - WWICS
  - RLG

OAI-PMH Details

- WHAT: A simple catalog record format and a harvester that can gather up the records or multiple sites
- WHY: To build richer services (browse and federated metadata search) across multiple sites
- HOW: Uses HTTP to transport XML records that contain Unqualified Dublin Core (DC) elements for ease of implementation

ISSUES

- Metadata inconsistency
- De-duplicating
- Restrictions
- Data Provider Participation
- Records for non-digital material
- Sustainability
- Technology bar set a little high for smaller institutions

Development

- Single domain services
- Heterogeneous services
- Level of granularity of the record – article or issue? guide or item? poem or collection?
- Access restrictions
- Maintenance
- OAI Static Repository

Metascholar Initiative

Martin Halbert

Emory University

Conversion Services (martin)

- Describe metadata conversion services; why you made the decisions you did and the rationale for those decisions.

Expert Scholars (martin)

- Describe the process of using expert scholars - specific examples of how they contributed.

Scholars input

- Give examples of how your project has incorporated feedback from the scholars

OAISTER

Chris Powell

University of Michigan

OAIster Statistics

- Online since June 2002
- Big
  - 1,183,995 records
  - 167 institutions (as of May 1, 2003)
- Popular
  - 11,101 search sessions
  - 32,878 searches (through Apr 30, 2003)
- Straightforward to implement and maintain

Three Issues Experienced by All Service Providers

- Metadata variation
- Access restrictions on digital objects described in OAI records
- Duplicate records for a single digital object

Metadata Variation

- As more records are available, users need more restrictions to focus in on what they want.
- Consistent metadata is needed to facilitate these restrictions.

Element Normalized

- **TYPE**: the obvious quick win
  - 240 native values mapped to four generic values (audio, video, image, text)

Audio = music, voice, etc.

Video = motion, animation, newsreels, etc.

Image = watercolour, watercolor, snapshot, slides, etc.

Text = article, articles, booklet, diss., story, etc.

Elements Not Currently Normalized

- **DATE**: where to begin?
  - 798,630 records with at least one date
  - records with up to seven dates included
  - no consistent style of date

Sample Date Values

```
<date>2-12-01</date>
<date>2002-01-01</date>
<date>0000-00-00</date>
<date>1822</date>
<date>between 1827 and 1833</date>
<date>18--?</date>
<date>November 13, 1947</date>
<date>SEP 1958</date>
<date>235 bce</date>
<date>Summer, 1948</date>
```

Elements Not Currently Normalized

- **SUBJECT**: out of context, what does it mean?
  - 547,738 records with at least one subject element
  - 99 records with 50 or more subjects
  - 1 record with 100 (Judson-Fairbanks Papers at CDL)

Sample Subject Values

<subject>30,51,52**</subject>**

<subject>1852, Apr. 22. E[veritt] Judson, letter to Philuta [Judson].**</subject>**

<subject>Slavery--United States--Controversial literature**</subject>**

<subject>view of interior with John Henry sculpture**</subject>**

<subject>Particles (Nuclear physics) -- Research.**</subject>**

Access Restrictions

- No records where the metadata itself is restricted in use (as far as we know!)
- Definitely some records where the objects are restricted to licensed users.

Access Restrictions (continued)

- DC **Rights** element doesn’t always give a sense of whether the item is available for viewing.
- Currently - no method of indicating restricted digital objects (i.e., OAI protocol “yes/no” toggle element).
- Assess whether users feel informed or frustrated by encountering restricted objects.

Duplicate Records for a Single Digital Object

- Acquired in two ways
  - Harvesting of originating repository and subject-specific aggregator
  - Harvesting of content host site and “static” OAI records provided by content creator

Aggregators

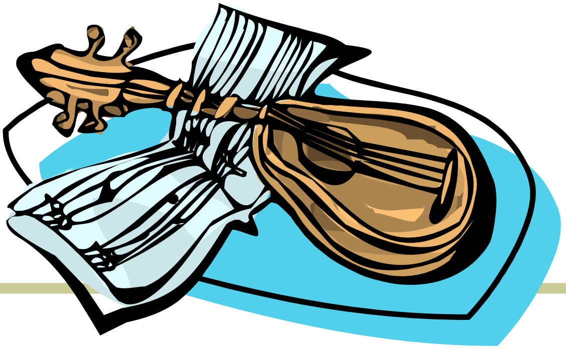
- Different identifiers for records, same object described and pointed to.
- Aggregators often contain records not currently available through OAI
- Aggregators do not always have all the records of a particular original repository – need to harvest both.

Content Host/Content Creator

- Records for objects provided twice
  - Repositories serving different roles for the same material.

- How to deal with the identified duplicates
  - Suppress?
  - Group?
  - Flag?

Sheet Music Harvester



Stephen Schwartz, PhD, MBA

Head of System Development for UCLA Library

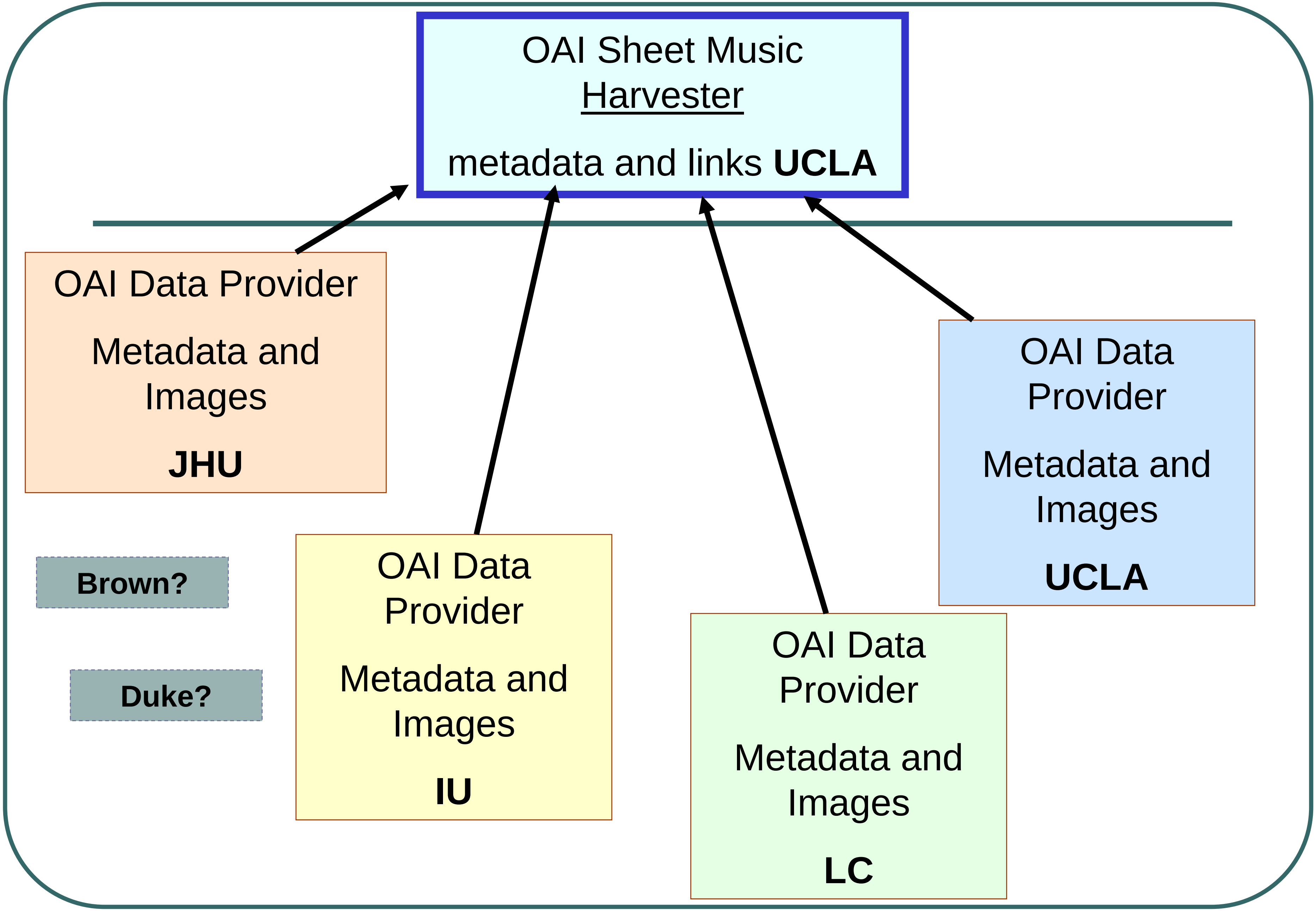
Some Observations

- Project Scope
- Best Practices for Data Providers
  - Data Mapping/Creation
  - Standards
- Researcher Usage Study Results with Corresponding Interface Design
  - Solutions

The Material

- The elements of a piece of sheet music
  - Text/lyrics
  - Music
  - Cover art
  - Advertisements etc.

- Sheet music as cultural resource



OAI Sheet Music Harvester

metadata and links **UCLA**

OAI Data Provider

Metadata and Images

JHU

Brown?

Duke?

OAI Data Provider

Metadata and Images

IU

OAI Data Provider

Metadata and Images

LC

OAI Data Provider

Metadata and Images

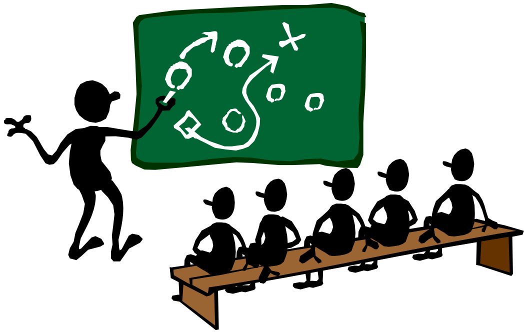
UCLA

Benefits of Specialized Harvester Collaboration

- **Improved Scholarship**
  - Enables more efficient research – saves time and travel
  - Enables new research – broader comparisons
- **Critical mass of Sheet Music data**
  - LC – 48,000 American Memory
  - IU – 1,500 of possible 100,000 Lilly Collection
  - JHU – 11,500 of possible 29,000 Levy Collection
  - UCLA – 1,100 of possible 450,000 American Popular
- **Shared expertise**
  - IU – Interface Design
  - JHU – Usability
  - UCLA – Harvester and Service Software and project coordination

Best Practice Guidelines

- Metadata Mapping/ Encoding Examples



  - Creator Roles
  - Dates
  - Descriptions and Lyrics
  - Added URL capability to DC Rights, Source, Relation fields

Unqualified Dublin Core

Identifier (URL)	Title	Creator
Contributor Not used.	Publisher	Date
Description	Format	Source
Type	Language	Subject
Relation	Coverage	Rights

DC_Creator

- “An entity primarily responsible for making the content of the resource. Examples of a Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.”
- **Composer, lyricist, arranger. Other contributors, including illustrator, artist, photographer, etc. Recommend last name first.**
- **Recommendation: Invert name. Use the authorized form of name where possible. If needed (e.g. for an alias) repeat the field for the alternative form.**
- Shaw, J. B., Jr. [lyricist]
- Stanley, J. Selwyn [composer]
- Lyrics and Music By Irving Berlin. Book by Harry B. Smith.

DC_Date

- “A date associated with an event in the life cycle of the resource. Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [W3CDTF] and follows the YYYY-MM-DD format.”
- **Date of publication. The most recent date to appear on the music, or, the actual date of publication if not present but known. Include other dates (e.g. date of composition) if known. Codes “c” for copyright and “ca.” for circa in front of the date is allowed for now. Use repeated DC fields for each date if needed.**
- 1920 [date of publication]
- 2002-07-17 [date of digitization]
- c1902
- [ca. 1800]
- 18--

DC_Description

- From Irving Berlin's Music box revue
- Lyrics: What a beautiful morning What a wonderful day We can hear the birds singing As we go on our way, Like a couple of children, We're so happy and gay for we've been married two years now And we're gonna be divorced to day. The judge is waiting at the court around the corner for the wife and me With the final decree You'll see a happy groom and bride Standing side by side Receiving congratulations When the knot is untied a foxy lawyer at the court around the corner will collect his fee the minute we're free We'll be divorced and then We'll soon be married again At the court around the corner the little wife and me
- Plate number: 6596-4
- Publisher number: 7466
- Respectfully dedicated to Miss Madeleine Cochrane.
- ads on back cover for Irving Berlin, Inc. stock

What did we learn?

Focus Groups

- Research needs
- Search strategies
- Desired content

Scenario Tests

- Inconsistent metadata
- Terminology
- Navigation
- Searching

Researcher Focus Group Results	Can Do
Keyword search	Yes
Browse by Name	Yes
Browse by role - composer, lyricist, arranger, or performer (etc.)	No
Date Range Browse	Yes
Browse by genre, subject, tempo, etc	No
Boolean Search and Browse	Yes
Save results – Music Stand	Yes
Share results with privacy controls – Virtual Collection	Yes

Some Researcher Tools

- **Music Stand**
  - working basket of selected music

- **Virtual Collection**
  - Private Use with notations
  - Controlled Sharing – e.g., course
  - Public Sharing

Search | Browse| Advanced Search | Music Box | Sign In | Virtual Collection

Welcome Stephen Schwartz Virtual Collections
Number of Items on Music Stand: **4**

	Collection	**Owner**	**Access**
My Virtual Collections			
	Jesus	shs	public
All Other Virtual Collections			
	Bananas	Stephen Davison	protected
	Guitar studies	Curtis Fornadley	protected
	Philadelphia Pub	Howard Batchelor	public

OAI Sheet Music Harvester - In Conclusion – after One Year

- OAI works:
  - low entry barrier
  - for specialized community/material
- Demonstrated institutional cooperation with mutual benefits:
  - Shared expertise - complementary
  - Variety of collections in size and metadata
- Received very encouraging feedback from researchers with simple DC
- Potential to develop some unique and new services – e.g., Virtual Collection

UIUC Digital Gateway to Cultural Heritage

Joanne Kaczmarek

Archivist for Electronic Records

University of Illinois Library

Background

- Project Scope
  - Developing software tools
  - Providing surrogate data provider services
  - Harvesting over 1M records – digital and non-digital

Archival Material

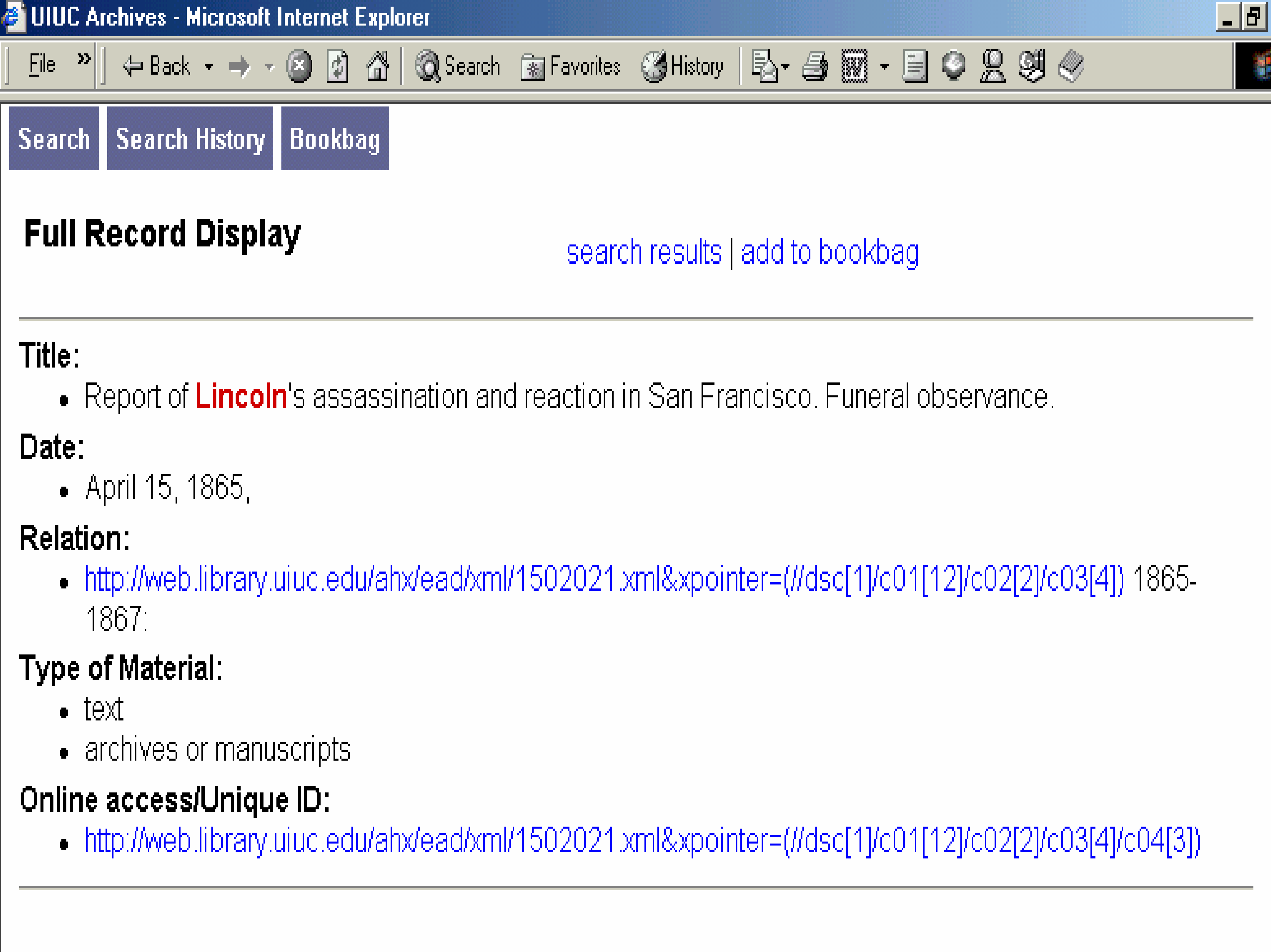
- Archival Description
  - The nature of archival material governs, rather than simply informing, the manner of description
- MARC-AMC, EAD, APPM
- Aggregated Services
  - Archives USA
  - Archival Resources

EAD Records

```
<archdesc type="inventory" level="collection">
<did id="a1"><unittitle>Irene Gomez-Bethke papers</unittitle><unitdate>1970-
1993.</unitdate></did>
  <c01>
    <did><unittitle>Hispanic Organizations in Minnesota:</unittitle></did>
      [other c02 levels not shown]
     <c02>
              <did><physloc>151.H.1.1B</physloc><container>1</container><unittitle>Archdiocesan
              Office of Hispanic Ministry:</unittitle></did>
       [other c03 levels not shown]
       <c03>
              <did><unittitle>Hispanic Ministry Advisory Board:</unittitle></did>
                            <scopecontent><p> Advised thearchbishop.</p></scopecontent>
          [other c04 levels not shown]
          <c04>
               <did><unittitle>Minutes of Board Meetings, 1986-1989.</unittitle></did>
          </c04>
        </c03>
      </c02>
   </c01>
</archdesc>
```

Preserving Context When Searching Item-level EAD Records

- Challenge – may inadvertently impede access by undermining provenance and original order
- Possible Solution – 2 XSLTs + XPointers
  - “top level” OAI record containing collection-level info
  - Description of subordinate components “dsc”
  - XPointer allows for identifying XML fragments



Full Record Display

search results | add to bookbag

Title:

- Report of **Lincoln**'s assassination and reaction in San Francisco. Funeral observance.

Date:

- April 15, 1865,

Relation:

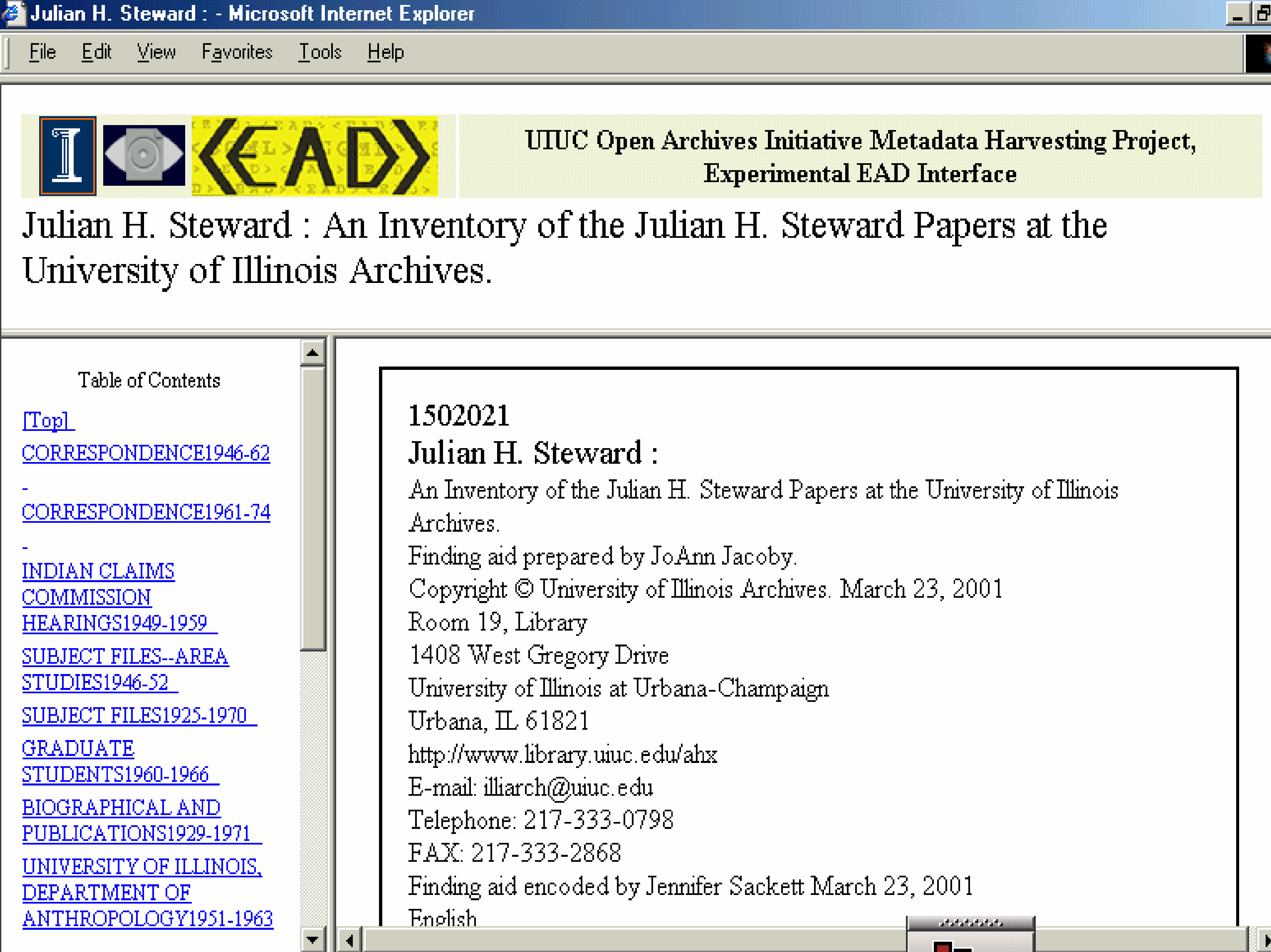
- http://web.library.uiuc.edu/ahx/ead/xml/1502021.xml&xpointer=(//dsc[1]/c01[12]/c02[2]/c03[4]) 1865-1867:

Type of Material:

- text
- archives or manuscripts

Online access/Unique ID:

- http://web.library.uiuc.edu/ahx/ead/xml/1502021.xml&xpointer=(//dsc[1]/c01[12]/c02[2]/c03[4]/c04[3])



UIUC Open Archives Initiative Metadata Harvesting Project,
Experimental EAD Interface

Julian H. Steward : An Inventory of the Julian H. Steward Papers at the University of Illinois Archives.

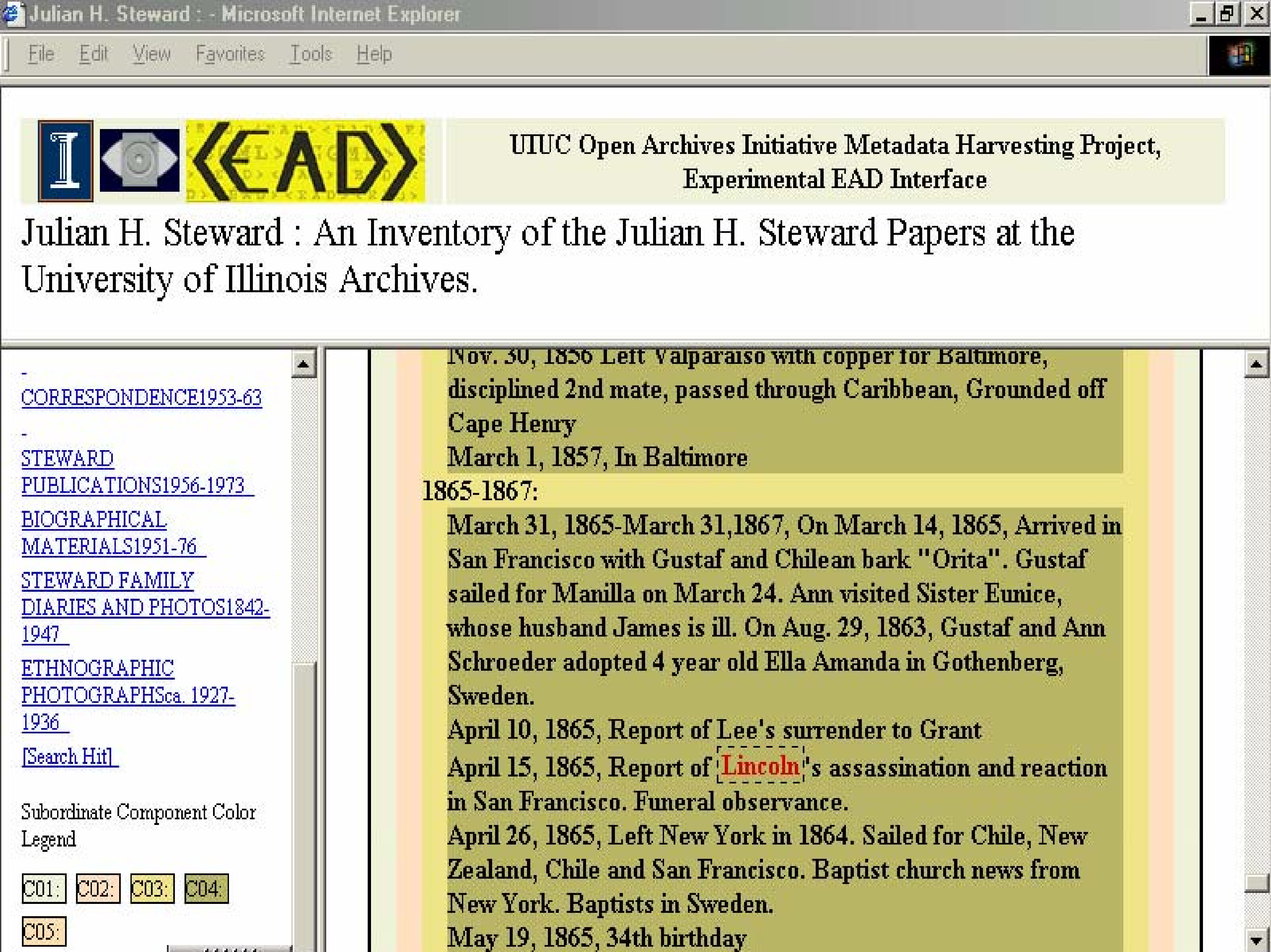
Table of Contents

[Top]
CORRESPONDENCE1946-62
-
CORRESPONDENCE1961-74
-
INDIAN CLAIMS COMMISSION HEARINGS1949-1959
SUBJECT FILES--AREA STUDIES1946-52
SUBJECT FILES1925-1970
GRADUATE STUDENTS1960-1966
BIOGRAPHICAL AND PUBLICATIONS1929-1971
UNIVERSITY OF ILLINOIS, DEPARTMENT OF ANTHROPOLOGY1951-1963

1502021

Julian H. Steward :

An Inventory of the Julian H. Steward Papers at the University of Illinois Archives.
Finding aid prepared by JoAnn Jacoby.
Copyright © University of Illinois Archives. March 23, 2001
Room 19, Library
1408 West Gregory Drive
University of Illinois at Urbana-Champaign
Urbana, IL 61821
http://www.library.uiuc.edu/ahx
E-mail: illiarch@uiuc.edu
Telephone: 217-333-0798
FAX: 217-333-2868
Finding aid encoded by Jennifer Sackett March 23, 2001
English



UIUC Open Archives Initiative Metadata Harvesting Project, Experimental EAD Interface

Julian H. Steward : An Inventory of the Julian H. Steward Papers at the University of Illinois Archives.

- CORRESPONDENCE1953-63
- STEWARD PUBLICATIONS1956-1973
BIOGRAPHICAL MATERIALS1951-76
STEWARD FAMILY DIARIES AND PHOTOS1842-1947
ETHNOGRAPHIC PHOTOGRAPHSca. 1927-1936
[Search Hit]

Subordinate Component Color Legend

C01: C02: C03: C04:
C05:

Nov. 30, 1856 Left Valparaiso with copper for Baltimore, disciplined 2nd mate, passed through Caribbean, Grounded off Cape Henry
March 1, 1857, In Baltimore

1865-1867:

March 31, 1865-March 31,1867, On March 14, 1865, Arrived in San Francisco with Gustaf and Chilean bark "Orita". Gustaf sailed for Manilla on March 24. Ann visited Sister Eunice, whose husband James is ill. On Aug. 29, 1863, Gustaf and Ann Schroeder adopted 4 year old Ella Amanda in Gothenberg, Sweden.
April 10, 1865, Report of Lee's surrender to Grant
April 15, 1865, Report of Lincoln's assassination and reaction in San Francisco. Funeral observance.
April 26, 1865, Left New York in 1864. Sailed for Chile, New Zealand, Chile and San Francisco. Baptist church news from New York. Baptists in Sweden.
May 19, 1865, 34th birthday

Ongoing Work

- There are about 100 OAI metadata providers currently registered on the OAI site and about a dozen OAI harvesting services. www.openarchives.org/Register/BrowseSites.pl There's also the Open Archives Forum site's even longer lists of European OAI sites www.oaforum.org/oaf_db/index.php.

- **IMLS – University of Illinois** – creating a registry of all digitized projects funded by Leadership Grants and providing support to data providers to become OAI-compliant

- **NATIONAL SCIENCE DIGITAL LIBRARY** – http://nsdl.org/ -- All NSDL contributors

OAI Ongoing Work

- **UIUC Grainger** http://g118.grainger.uiuc.edu/engroai/ An OAI search service for Engineering, Computer Science and Physics

- **Virginia Tech** ImageBase project and other OAI projects (http://imagebase.lib.vt.edu/about.php)

- **The Open Language Archives** http://www.language-archives.org/ Probably the best example of a broad community-based OAI metadata harvesting service (http://linguist.emich.edu/olac/).

OAI Ongoing Work

- ***OCLC* --** http://alcme.oclc.org/index.html Top level page for a number of OCLC projects related to aggregation, including an OAI metadata provider for all 4 million+ Theses and Dissertation records in WorldCat
- **LC American Memory** project **--** http://memory.loc.gov/ A large portion of those collections are also available via OAI for harvesting.
- **Old Dominion University's ARC --** http://arc.cs.odu.edu/ The original OAI harvesting service.

OAI Ongoing Work

- **Southampton's EPrint service** An extensive OAI search service http://citebase.eprints.org/cgi-bin/search
- **Colorado Digitization Program** An important IMLS collaboration project with metadata now available via OAI. The CDP model is being expanded for the multi-state Western Trails interoperability project http://www.cdpheritage.org/
- **DLF** – OAI records coming out for DLF publications; tester of the static repository

Miscellany

- Open Source OAI tools: SourceForge
- Growing community of practice and strong endorsement
- Some hope of persuading vendors to provide OAI for journal articles
- Clearly meeting a core need
- Sustainability
- Relevance to ACH/ALLC?