



The Open Archives Initiative and the Sheet Music Consortium

Jon Dunn, Jenn Riley
IU Digital Library Program

October 10, 2003



Presentation outline

- Jon:
 - OAI introduction
 - Sheet Music Consortium background
- Jenn:
 - Data mapping issues
 - Sheet music harvester demonstration
 - Next steps

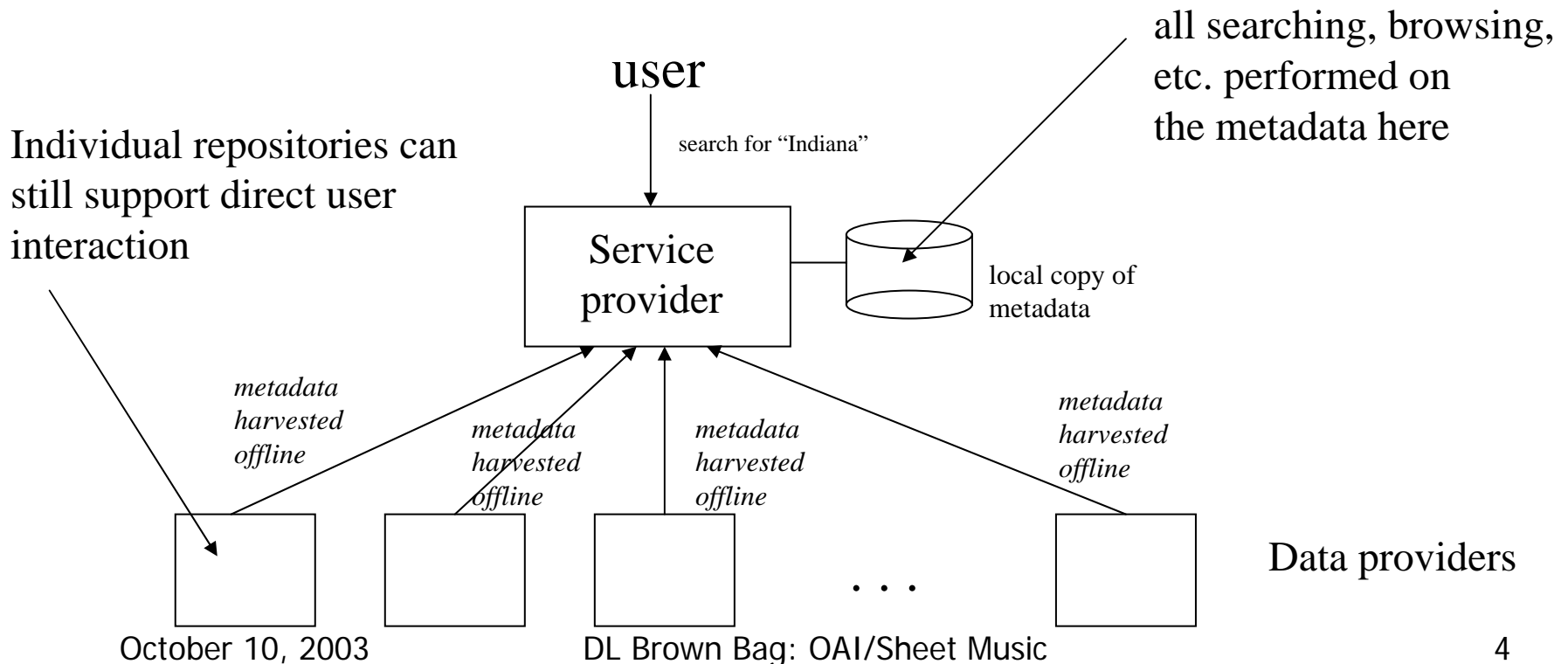


OAI: Open Archives Initiative

- Original problem: searching across e-print archives
- Distributed searching hard
 - e.g. Z39.50
 - Varying search semantics, capabilities
 - Network, server problems
- Solution: metadata harvesting
 - OAI-PMH: OAI Protocol for Metadata Harvesting

Metadata Harvesting

- Extract metadata from various sources
- Build services on local copies of metadata





OAI-PMH roles

- Data Providers
 - Repositories of digital content and metadata
 - Support harvesting of metadata via the OAI protocol
- Service Providers
 - Harvest metadata from data providers using the OAI protocol
 - Implement user interface to data
 - Usually for searching, but other services also possible
 - Can be selective



OAI Protocol for Metadata Harvesting

- Originally developed in 1999 (Santa Fe Convention)
- Original focus on E-prints
- Has grown into general metadata harvesting protocol
- Version 1.0: January 2001
- Version 1.1: June 2001
 - Conform to XML Schema 1.0
- Version 2.0: June 2002
 - Transition period through December 2002
- Currently 120 *registered* OAI data providers (up from 53 in March 2003)



OAI-PMH tech details

- Carried over HTTP
- Requests: HTTP GET or POST
- Responses encoded in XML
 - Format defined via XML schema
- Metadata in unqualified Dublin Core (and potentially other formats)



Dublin Core elements

- Coverage
- Description
- Type
- Relation
- Source
- Subject
- Title
- Contributor
- Creator
- Publisher
- Rights
- Date
- Format
- Identifier
- Language



OAI-PMH verbs

Verb	Function
Identify	description of archive
ListMetadataFormats	metadata formats supported by archive
ListSets	sets defined by archive
ListIdentifiers	OAI unique ids contained in archive
ListRecords	listing of N records
GetRecord	listing of a single record



OAI resources

- Web site, mailing lists
- Repository explorer
- Data/service provider software

www.openarchives.org



OAI data providers at IU

- OAI data provider for DLP collections
 - Lilly: Hohenberger Photograph Collection, DeVincent Sheet Music Collection
 - IUN: U.S. Steel Photograph Collection
 - eventually all
- Eprints: Digital Library of the Commons
- AISRI
- ReciprocalNet



OAI data provider for DLP

- PHP OAI Data Provider
 - Developed by University of Oldenburg
 - PHP, MySQL database
- Perl scripts used to map USMARC, other formats to DC
 - MARC.pm Perl module



Examples of OAI service providers

- UIUC Digital Gateway to Cultural Heritage Materials
 - <http://oai.grainger.uiuc.edu/>
- UMich OAIster
 - <http://www.oaister.org/>
- RLG Cultural Materials (licensed)
 - <http://www.rlg.org/culturalres/>
- OLAC: Open Language Archives Community
 - <http://www.language-archives.org/>



Sheet Music Consortium

- Partners
 - UCLA
 - Johns Hopkins
 - IU
- Goal: Integrate access to sheet music collections
 - Online and print collections



Sheet music

- Definition
 - Based on physical format: generally loose sheets or folio, 1-10 pages
 - Much is “popular music,” but not all
- Variety of research uses
- Currently hard to access
 - Variety of metadata
 - Much uncataloged
 - Many valuable collections
 - [MLA list](#)
 - At IU: [Lilly](#), [Archives of Traditional Music](#)

THE GHOST OF THE UKULELE

BY
JAMES BROCKMAN
JACK SMITH



THE HIT OF
Shubert's Production
"SHOW of WONDERS"

SUNG BY
HOWARD & HOWARD

AT THE
WINTER GARDEN
NEW YORK CITY

JAMES BROCKMAN MUSIC PUB. CO.
145 WEST 45TH STREET
NEW YORK.



The Ghost Of The Ukulele

By JAMES BROCKMAN
and JACK SMITH

Allegretto Moderato

Piano

Till ready

The piano introduction consists of two staves. The right hand plays a series of chords and eighth notes, while the left hand provides a steady bass line. The tempo is marked 'Allegretto Moderato' and the dynamics include 'f' and 'p'.

Last night in dreams I could plain-ly see, The land of Hon-o-lu-lu far a-cross the sea... I was
When I a-woke I was full of fright, From hearing u-ku-le-le mus-ic played all night, - I

The first system shows the vocal line and piano accompaniment. The vocal line is on a single staff, and the piano accompaniment is on two staves. The lyrics are: "Last night in dreams I could plain-ly see, The land of Hon-o-lu-lu far a-cross the sea... I was When I a-woke I was full of fright, From hearing u-ku-le-le mus-ic played all night, - I".

on the beach at Wai-ki-ki when a mil-lion u-ku-le-les start-ed chas-ing me, - Bing - Bing -
locked the door, turned out the light, crawled back in-to my bed and hid right out of sight - Bing - Bing -

The second system continues the vocal line and piano accompaniment. The lyrics are: "on the beach at Wai-ki-ki when a mil-lion u-ku-le-les start-ed chas-ing me, - Bing - Bing - locked the door, turned out the light, crawled back in-to my bed and hid right out of sight - Bing - Bing -".

That's the sound that's haunting me - Bing - Bing - The u-ku-le-le mel-o-dy -
That's that mel-o-dy a-gain - Bing - Bing - That sneaky creepy fun-ny strain -

The third system continues the vocal line and piano accompaniment. The lyrics are: "That's the sound that's haunting me - Bing - Bing - The u-ku-le-le mel-o-dy - That's that mel-o-dy a-gain - Bing - Bing - That sneaky creepy fun-ny strain -".

Chorus

That's the ghost, that's the ghost, that's the ghost of the U-ku-le-le - Strumming gay-ly - How they

The fourth system shows the chorus. The vocal line is on a single staff, and the piano accompaniment is on two staves. The lyrics are: "That's the ghost, that's the ghost, that's the ghost of the U-ku-le-le - Strumming gay-ly - How they". The piano part includes dynamics like 'mf' and 'f'.

sway when they play on the gay lit-tle u - ku - le - le - play-ing dai - ly

First it makes you sad Then it makes you glad Take it a - way I say Take it a -

way I say take it a - way cause its driv-ing me mad That's the ghost that's the ghost that's the ghost of the u - ku -

le - le - Hear it da - i - ly - no mat - ter where I go it fol - lows me *(There it is, there it*

is there it is) My dig-ni - ty I've tried to keep - But I did the hu - la hu - la in my sleep - From the

ghost of the u - ku - le - le mel - o - dy That's the *ad lib* dy.

TRY THIS OVER ON YOUR PIANO

Why I Love You, I Don't Know.

JAMES BROCKMAN
and
JACK SMITH

Chorus.

I know why the sea is nev-er still, I know why the tem-pests blow,
I know why in spring all the birds sweet-ly sing, I know why the flow-ers grow,
I know why the Moon is a-bove the clouds, I know why the Sun sets low,
know in my heart, I've been fuled from the start, But why I love you, I don't know.

Copyright MCMXXII by James Brockman Music Pub. Co. 146 W. 45th St., N.Y.C.
Mechanical Rights Reserved. International Copyright Secured.

FOR SALE BY ALL DEALERS!

L-117
M1
1548
by 271
no 44
copy 1



Sheet Music Consortium

Harvester: Timeline

- March 2002: Initial planning meeting at IU
- Fall 2002: Initial system prototype
- Winter 2002/2003: Usability evaluation, interface redesign
 - Focus groups and usability testing at several sites
- Fall 2003 – Version 1 of system released



Why did we have to map data?

- OAI requires unqualified Dublin Core
- Sheet Music Harvester version 1 only collected Dublin Core
- Contributed data only needed to support resource discovery
- Dublin Core field definitions need interpretation
- For efficient searching, data from different institutions must be consistent



Some mapping issues

- Field formatting important, not just contents
- Choices heavily influenced by LC practice
- Can't force institutions to comply with guidelines
- Sheet music has many alternative titles
- Creator vs. contributor
- Plate numbers: they're important, where to put and how to label?
- Uncertain dates and date ranges



Mapping guidelines

- **Examples:**
 - **Creator:** Invert name. Use the authorized form of name where possible. If needed (e.g. for an alias) repeat the field for the alternative form.
 - **Date:** 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.



Existing metadata formats

- MARC
- Encoded Archival Description (EAD)
- Dublin Core (DC)
- Local custom formats



MARC (1)

- Library of Congress – mostly from Music for the Nation: American Sheet Music, 1820-1860 & 1870-1885
 - almost 50,000 records available via OAI
 - already had data mapped “based on” MARC to Dublin Core crosswalk
 - not able to alter their mapping for participation in sheet music project



MARC (2)

- IU – Starr collection
 - little authority control
 - determined LC MARC2DC mapping inadequate
 - mapping in progress using MARC.pm
- Duke – Weinmann collection
 - rare materials emphasis
 - also customized own mapping
 - mapping in progress



EAD

- Duke – Historic American Sheet Music
- Item level finding aid
 - very robust and specific
 - conversion was relatively simple because data was converted to EAD from collection-specific database
 - included virtually all information in EAD documents to DC records



Dublin Core

- UCLA – [Archive of Popular American Music](#)
- 4 types of DC records
 - songs
 - sheet music
 - covers et al
 - recordings
 - mapping only required inheritance of songs and sheet music data elements down to the covers level
 - recordings data ignored for OAI data provider purposes



Local custom formats (1)

- Johns Hopkins – [Levy collection](#)
- Simple SGML DTD
 - publication (location, publisher, date)
 - subject
 - call num (box, item)
 - title
 - composer/lyricist/arranger
 - form of composition
 - instrumentation
 - first line
 - first line of chorus
 - performer
 - dedicatee
 - engraver/lithographer/artist
 - advertisement
 - plate num
 - duplication



Local custom formats (2)

- IU – DeVincent collection
- Simple MS Access database
- Conversion done with Perl
 - title
 - composer
 - lyricist
 - place of publication
 - publisher
 - copyright
 - first line
 - first line of chorus
 - subject
 - form of composition
 - performance medium
 - copies
 - call #



Harvester demonstration

- <<http://digital.library.ucla.edu/sheetmusic>>



Data inconsistencies

- Different depths of description
- Different levels of authority control
- No common subject vocabulary between collections
- Despite mapping guidelines, differences in DC interpretation



Next steps?

- Authority control for names
- Date formats
- Data clean-up: what can be done at harvester end and what must we ask data providers to do?
- What will more robust data format look like?
- How do we make it easier for more institutions to participate?



More information

- Presentation on DLP web site, with links:
 - www.dlib.indiana.edu/workshops/bbfall2003.htm
- Email:
 - Jon Dunn: jwd@indiana.edu
 - Jenn Riley: jenlrile@indiana.edu