

AIRS Digital Library (AIRS DL)

Use Case Report: A Digital Library for Research on the Singing Voice

23 Feb 2012, Draft 1

High-Level Goals of the AIRS Digital Library

The AIRS Digital Library project will foster the sharing of data and findings among researchers of the singing voice. AIRS researchers are compiling a unique set of resources, and as a result, they have atypical sharing and archiving needs.

Each of the AIRS research themes points to several disciplines. For example,
(This list is not exhaustive, but represents the diversity for the purposes of designing a data archive.)

- **Development** – developmental psychology, perception and cognitive science, neuroscience, linguistics and speech
- **Education** – various branches of psychology, social psychology, education relating to pedagogy and epistemology
- **Well-being** – medicine, physiology, sociology and neuroscience

This diversity is to be represented by the AIRS DL. Also, the needs of AIRS researchers, as individuals and as a collective, will be addressed through the AIRS DL's design and functionality.

The Library will adopt a “vocabulary” that holds specific meaningful for singing voice research and that will encourage collaboration among disciplines.

The Library will persist even after the AIRS project has ended. Ideally, the contents and the design will serve researchers of singing voice beyond AIRS, and potentially it may offer resources for singers outside of a research context.

The design of the AIRS DL will capture all of these facets.

AIRS Digital Library as a Resource

The AIRS DL will need to house a large variety of resource types including

- audio/visual data
- corpora (comprised of the above)
- tabular data sets
- media files and numerous publications for reference purposes
- pedagogical materials

It will encourage innovative approaches to research primarily through its structure. Various types of resource materials, raw data and published studies will be interconnected through consistent descriptions, metadata and search functions. In this way, studies, findings and data sets will be made meaningful across disciplines.

Moreover, it is our intention that data will be tagged with the methodologies that have been used to analyze it. In this way, we will facilitate researchers finding new data suitable to their approach, and also encourage researchers to consider new approaches.

Scope of the design challenge

The macro-level design aims include

- A specialized singing research vocabulary to form the basis of an intuitive system of

- (meta)tagging and descriptions.
- Cross-referenced content including data, media examples and other resources (musical scores, lesson plans, perhaps computer code)
- Enable data set sharing

Design Methodology

The AIRS Digital Library design phase consists of several stages: data collection, analysis and synthesis. (Some of these steps have already begun and are on-going.)

- Identify the library needs of AIRS researchers (for sharing and archiving)
- Assess the data types and content that has been collected through AIRS
- Define the content models (AIRS DL will house various formats)
- Assess how to identify that content (i.e. how researchers/users tag it, search it, refer to it for their own purposes)
- From that develop a specialized “vocabulary”
- Determine how to organize the content in to collections (content models)
- Collections may house “series” of related data. A collection, may have several distinct series in a collection.
- Determine the library’s architecture, in the case of AIRS, to be built on Islandora. (This will potentially require developing custom content models.)
- Customize collection forms and ingest process to capture the necessary identify data (as determined by the vocabulary.)
- Customize search functionality (according to vocabulary and content models)

Best Practices

The Data Documentation Initiative (DDI) [1] has developed a metadata specification for the social and behavioral sciences, and moreover, best practices for data management. Many academic libraries including MIT recommend these practices. (Robertson Library’s VRE information page points to MIT’s recommendations). DDI’s best practices include recommendations [1] for beginning a new data management project. These steps are outlined below.

This document from here on out adopts wherever appropriate DDI’s language and terminology. The steps listed below are DDI’s language (direct quotes). The descriptions beneath, except where noted, relate DDI’s concepts to the AIR DL.

Start with the Study Unit

DDI calls the study unit “an instance”. It is essentially a “citation” for the content. They recommend at this stage recording the purpose of the study and the concepts pertinent to the study.

Move on to Data Collection

Questionnaires are the recommended method for collecting (textual) information about the researchers’ data including (the methodology, collection event, and data processing.)

Design the Logical Product (variables section)

The data product is the result of codifying the results from the data collection stage. The data product represents the variables necessary to describe each study unit contained in the Digital library.

Create a Physical Data Product

“The Physical Data Product describes the physical structure of the data store.” For Islandora, this will connect to the content models.

Create the Physical Instance

The Physical Instance description bears a one-to-one relationship of a physical instance to a data file. We will need an appropriate content model for all that we ingest.

Pull it all together

Each single Study Unit is ultimately ingested (in accordance with the appropriate content model).

Make a group

Common information connects study units. They may be organized into groups, series or “collections” in Islandora. In Islandora there are parent and children collections.

The UK Data Archive is another major archiving initiative that offers additional data management best practices and resources. In their “Best Practices Guide” [2] they note that “[w]henver data are used sufficient contextual information is required to make sense of that data.” They suggest that data-level documentation include the following [direct quote]:

For the dataset

- the context of data collection: project history, aim, objectives and hypotheses
- data collection methods: sampling, data collection process, instruments used, hardware and software used, scale and resolution, temporal and geographic coverage and secondary data sources used
- dataset structure of data files, study cases, relationships between files
- data validation, checking, proofing, cleaning and quality assurance procedures carried out
- changes made to data over time since their original creation and identification of different versions of data files
- information on access and use conditions or data confidentiality

For each data element

- names, labels and descriptions for variables, records and their values
- explanation or definition of codes and classification schemes used
- definitions of specialist terminology or acronyms used
- codes of, and reasons for, missing values
- derived data created after collection, with code, algorithm or command file
- weighting and grossing variables created
- data listing of annotations for cases, individuals or items

AIRS Data Management

Through the AIRS DL initiative we have already started the process recommended by DDI with the intention of collecting the necessary vocabulary and variables to create an AIRS singing voice library.

Recently, a survey went out to ~ 25 principle investigators. This first survey will provide a basis for the vocabulary and help us to determine the kinds of data sets created through AIRS and the contextual information that will be needed to define a “study unit” in the AIRS DL.

It must be noted that the AIRS project has been underway for years. Data has already been collected and this places some constraints on the AIRS DL design. We cannot undo/redo existing data. Also, our choice of platform, Islandora, may necessitate limiting the content types in the archive and perhaps also the content and format of the contextual data attached to each study unit.

File formats for sharing, re-use and preservation

The following draws from several sources including the UK Archive’s guideline [2]. The file types listed are included as examples and should not be considered recommendations at this stage of

development.

For quantitative data, tabular dataset are common, and many formats offer the option of including extensive metadata such as variable labels, code labels and defined missing values. This information is packaged with the matrix of data. SPSS, Stata and SAS are common files and support these features. Quantitative data in tabular formats and minimal metadata are also common as can be found in comma-separated values (CSV) files (.csv) or tab-delimited files (.tab). Some of the file types listed above may have proprietary issues, but they are also ubiquitous. *Many archivist caution against proprietary formats.*

Qualitative data captured in textual form may include eXtensible Mark-up Language (XML) text, but Rich Text Format (.rtf) and plain text data, ASCII (.txt) are also appropriate archival formats.

There are numerous proprietary formats for audio and visual files, and they are ideally to be avoided. JPEG is an option for image data. For digital audio data Free Lossless Audio Codec (FLAC) (.flac) offers an uncompressed option and MPEG could offer a compressed option. Similarly for digital video data there is MPEG-4 (.mp4).

For documentation there are Rich Text Format (.rtf), PDF (.pdf) and OpenDocument Text (.odt) formats.

Incorporating Existing Data Sets

AIRS has options for incorporating existing data sets in less than desirable file formats. Depending on the formats in question and the size of the data set, it may be feasible to *transcode* from one file format to another without a loss in quality (or data). Alternatively, if the data itself cannot be ingested into the digital library, hyperlinks may be embedded into catalog descriptions that point the user to storage locations outside the library. The catalog descriptions would be searchable within the library, even if not all search functions were available.

Not all existing data will fit neatly into the long-term AIRS DL solution. The AIR DL can improve the less-ideal alternatives for archiving orphan data sets by *capturing information about the data now while data is still being collected* (and institutional memory is fresh). AIRS DL can also develop a set of *best practices moving forward* that is specific to AIRS for collecting data pertaining to singing voice. These steps will allow AIRS to optimize the DL design to house what AIRS has now and what AIRS will have in future.

Use Cases

What is the best design for the AIRS DL? Several use cases have been deconstructed in terms of their functionality relative to the nature of the content, the research community using the resources, and how the content is accessed.

The use cases described below are not all examples of singing voice repositories. Not all deal with music or even sound (although most of them do). These cases were chosen to represent a wide spectrum of audiences, content and design strategies. They have been categorized by size and the diversity of the content. Several of these examples try to serve several intended audiences simultaneously - including experts and musical professionals, academics and the general public.

Categories

Institutional or Idiomatic archives: These archives are geared towards the preservation of the historical records of an individual (renowned) institution or an art form/idiom. They might in some cases be considered “vanity” or “fan” archives, but tend to house primary sources that are valuable to

academic researchers and unavailable elsewhere. The institutions are themselves “primary sources”. In some cases, these are open-ended rather than legacy repositories, and new content is still being added. The opera archive at the Teatro alla Scalla is a prime example .

Historical Records: These archives form a kind of “public” record and must serve both academic experts and the general public. These tend to be very large collections or collections of themed collections or series. New collections are added more often than new pieces of content are added to an existing collection. The US Library of Congress exemplifies this. Harvard’s Milman Parry Collection of Oral Literature and South Slavic Heroic Song is another example. Although much, much smaller in size, it is designed to house the collected materials of individual researchers. Each researcher’s collection is essentially closed/complete.

In some cases, these archives are organized by content type, with little structure provided to organize the materials into series or themed collections. The Irish Traditional Music Archive follows this approach, and offers only searches by content type.

Small, Specialized Collections: These archives tend to set a singular curatorial message. For example, the University of Southern California’s Gospel Music History Archive strongly communicates the historical significance of the content contained within.

Search by Subject, Limited Content Types: Due to the subject matter, some archives are able to restrict the content types and search strategies. The Macaulay Library at Cornell University houses a vast amount of bioacoustics samples. All the content is preserved in one audio or one video file format. In this case the “subject” is animal species.

Shared Data Sets For Narrow Areas of Study: Some disciplines share data sets for benchmarking tasks (for example to test different types of algorithms or iterative algorithm refinements on the same task). The data sets are in many cases geared towards one mode of analysis or intended for use in testing a particular type of tasks. The Music Information Retrieval community shares several such data sets.

Repositories for Sharing Data Sets: Several large-scale projects have been undertaken in the social sciences to archive vast amounts of data and make it available to all researchers. All the structure and functionality of these repositories is geared towards helping researchers find data sets they can use and preserving data sets they have created.

Archives of Teaching Materials: Some library house atypical materials, for example, teaching materials. These resources require special categories and descriptions, as can be seen in Stanford’s Language Materials Project.

Cases by Type

Institutional or Idiomatic archives:

Teatro alla Scalla
<http://www.archiviolascala.org/>

The opera archive at the Teatro alla Scalla is not designed for academic rigor. Its design first and foremost promotes the continued historical significance of the Teatro alla Scalla in opera. The archive is part of an outreach strategy. It serves enthusiasts. However, the materials it houses hold interest for researchers, especially since these are primary sources provided by the primary source.

From the theater’s homepage, there is a top-level link to the archive. On the archive page the search criteria are listed each has a dedicated text box.

Search criteria:

- Title
- Season
- Composer
- Performer
- Character
- Place

This design maps 1:1 the archive's vocabulary and the study units' descriptions. For non-experts it is easy to use, and it encourages exploration by these criteria.

Historical Records:

<http://www.loc.gov/rr/record/onlinecollections.html>
<http://chs119.chs.harvard.edu/mpc/index.html>
<http://www.itma.ie/digitalibrary/playlist/medley-1950s/>

US Library of Congress

<http://www.loc.gov/rr/record/onlinecollections.html>

The Library of Congress collection is enormous. It is a collection of collections, and there is little consistency among the "children" collections contained within. Some collections are geared more toward scholars while others have a very developed public outreach component.

The LOC's holdings expand often with the addition of new, themed child-collections. Nonetheless, the impression is that each individual child-collection is static. It is often hard to search across child-collections. While this is perhaps not a major concern for academic researchers who are adept at searching, the site does not facilitate new research. Researchers do not interact with other researchers. There are no pointers indicating what happens to this information.

The Audio Collections landing page is under the Motion Picture, Broadcast & Recorded Sound Division.

The LOC >> Researchers >> Recorded Sound Reference Center >> Online Audio Collections and Presentations

The content itself is quite deep.

The children-collections are categorized sometimes, but not always, by the study center that produced the child-collection. Study centers often but not always generate series for their child-collections. On main landing pages, categories and child-collections are organized alphabetically. Clicking on any child-collection links out to a child-collections homepage. Each child-collection has its own organizational hierarchy and search functionality. For example, the National Jukebox offers search by

- Artist
- Genre
- Year
- Playlists
- Themed collections

The user needs to be knowledgeable about idioms to make use of it. At a deeper level, there are advanced searches that map to the study units' vocabulary.

Milman Parry Collection of Oral Literature Repository of South Slavic Heroic Song
<http://chs119.chs.harvard.edu/mpc/index.html>

The Parry Collection is drastically smaller in size than the LOC, but takes a similar approach and shares many of the same pros/cons. Search options include

- Search for keyword (in singer/title/date/place/cat number)
- Collection method (copied/recorded/dictated/autograph)
- Collection
- Digital audio (check box)
- Digital text (check box)

And

- Browse by-> (singer/title/date/place/)

Like the LOC, the curatorial message is neutral. It is geared toward users with academic research skills, but it does not dissuade enthusiasts. The user needs to know what they want. They need prior familiarity with the materials, even if they “browse”. Unlike the LOC, it contains much fewer content types, and it makes the search vocabulary (criteria) apparent at the top-level of the search page.

The Irish Traditional Music Archive
<http://www.itma.ie/digitallibrary/playlist/medley-1950s/>

The Irish Traditional Music Archive is vast. A large collection with few categories, it offers a wealth of resources— that is the curatorial message— but it is hard to search. Although its is likely to have what is needed, finding the required content may seem difficult, unless the research has done a lot of research about it before the search. Search options include

- Playlists (themed series)
- Sound Recordings
- Video Playlists
- Videos
- Galleries (photo series)
- Images
- Printed collections (document series)
- Printed Items
- Interactive Scores (pages turn automatically)

The hierarchy suggests that the designers are expecting users to browse. Under these subheadings, are alphabetized lists and search by keyword

--- Keywords
 general
 in title and entries
 exact phrase match

The interactive score feature is a beautiful teaching utility and significantly enhances the relevance of historical and archived materials for practicing musicians. Through this feature and also the breadth of the geographic locations represented, this archive succeeds in presenting the idea that this is a “living” collection. Interestingly, part of what communicates this is that there is no dedicated search by location function— although one can search by a location keyword.

Small, Specialized Collections:

<http://digitallibrary.usc.edu/gmha/controller/browse.htm?c=2&t=People&summary=ROLE&mode=list>

<http://www.globalmusicarchive.org/about.html>

University of Southern California's Gospel Music History Archive

<http://digitallibrary.usc.edu/gmha/controller/browse.htm?c=2&t=People&summary=ROLE&mode=list>

The University of Southern California's Gospel Music History Archive has limited holdings. While individual study units do have value for a variety of research goals, taken as whole the archive's components paint a picture of the evolution of gospel. Limited in size and scope as it is, researchers are not likely to go there unless it has what they need.

It is easy to search. The top level offers

- Interactive timeline

- Browse by
 - Style/Genre
 - Spirituals
 - Early Church Music
 - Early Gospel Blues
 - Trad Gospel
 - Quartet
 - Contemporary
 - Urban/Hip-hop

- People
 - Singers (Soloist/Groups/Choirs)
 - Composer/Arranger
 - Lyricist
 - Gospel Preacher
 - Musicians
 - Promoters/Managers
 - Broadcasters

- Time period
 - Roughly 40 year intervals to the present

- Geography
 - (US) regions
 - Cities

Many of these categories are empty. The archive houses various formats and study units are subject to differing permission policies. In restricted use cases, users do not access the files directly. However, the data information itself is valuable for researchers. Contact info for permissions, formats, etc. available with each study unit.

Global Music Archive at Vanderbilt
<http://www.globalmusicarchive.org/about.html>

The Global Music Archive at Vanderbilt is another academically-oriented, selective archive. It focuses on the African Diaspora. Recordings (at present limited to East African), documents, bibliographies and cassettes each have their own top-level menu. To search the recordings, users navigate

Collection>> Browse by Category

- Artist/Group
- Region
- District
- Language
- Ethnic Group
- Musical Instrument

Basic (keyword) Search

The advanced search offers several options mapped to study unit vocabulary

General

- Keyword
- Title
- Group/Emblem
- Performer

Location

- Performance venue
- City/village
- District
- Region

Culture

- Language
- Instrument
- Ethnic Group

Search by Subject, Limited Content Types:

The Macaulay Library, Department of Ornithology, Cornell University
<http://macaulaylibrary.org/>

The Macaulay Library houses a vast and expanding archive of bioacoustics data. It restricts its collection to audio and video files only (and limited formats). "Search by Species" is the primary mode of access. However, browsing is possible.

Browse by >>

Taxonomy >> Animalia >> (Chordata*) >> (Aves) >> (Sphenisciformes) >> (Spheniscidae)
(* all species have separate categories)

The results are returned by

Audio | Video by Spheniscidae - Animal (in English i.e. Penguins)

Results may be sorted by catalog, species, location, date, recordist, length or quality.

Although designed to house and make accessible scientific data, it is very user-friendly. Browsing is fun (and enhanced by the animal photos). There are several “staff pick” collections available. However, to find specific sounds, the user needs to know the species.

Shared Data Sets For Narrow Areas of Study:

<http://staff.aist.go.jp/m.goto/RWC-MDB/>
http://ismir2004.ismir.net/melody_contest/results.html#_Toc84660183

Real World Computing database
<http://staff.aist.go.jp/m.goto/RWC-MDB/>

Some research communities rely heavily of shared datasets. For example, in Music Information Retrieval, researchers developing algorithms commonly share datasets so that they may compare algorithms. They may benchmark and test the performance of an algorithm over iterations of refinements, and/or they may compare wholly different algorithms’ performances on the same test set.

Such repositories are intended to facilitate the sharing of data only. Researchers will not search these databases for particular study units.

Real World Computing database is copyright cleared, and is available as a common foundation for research. Permission to access the database must be granted by the governing body. It is comprised of the following:

- Popular music (100 songs)
- Classical Music (50 pieces)
- Jazz Music (50 pieces)
- Genre database (contains 10 genre categories – for use in identifying genre research)
- Instrument sounds (50 instruments)

ISMIR Audio Description DataSet
http://ismir2004.ismir.net/melody_contest/results.html#_Toc84660183

The Music Information Retrieval Exchange (MIREX) administers an annual contest to evaluate music information retrieval algorithms. MIREX works in conjunction with the International Society for Music Information Retrieval (ISMIR). MIREX provides test sets for the algorithms entered into their annual competition. The 2004 set became “canonical” for a while. MIREX makes the sets available after the competition whenever possible so that comparisons may be made against newer algorithms. The 2004 test set includes, 10 new audio excerpts (for the competition) and 10 audio excerpts from the tuning set.

Repositories for Sharing Data Sets:

Inter-University Consortia for Political and Social Research
<http://www.icpsr.umich.edu/icpsrweb/content/ICPSR/access/deposit/guide/>

The Inter-University Consortia for Political and Social Research’s sole purpose is to archive and facilitated the sharing of data sets in political and social science research. The method of analysis applied to the data is a primary identifier for each study unit. In this way, the archive helps researchers find data appropriate for their mode of analysis and also offers insight into the kinds of results that have been produced using the data.

This repository has varied types of data sets and varied ways to search data sets. It also offers organized themed series.

The main search menu offers

- Find and Analyze data >>
- Find Data >>
- View all studies >>
- View all studies for which online analysis is available >>
- Browse by topic >>
- Browse by geography (interactive map, by country) >>
- Browse by investigator >>
- Browse by Series >>
- Browse by recent updates and Additions >>

Under “Find and Analyze Data” >>

- Variables database
- List all studies
- List all series

There is an advanced search (queried by “variable label, question or answer”)

Under “Find and Analyze Data” >>

- Analyze online - There is the choice of limiting results to data that can be analyzed online (using interactive functions).
- Find and Analyze data >> Theme collections
- Find and Analyze data >> Restricted data
- Find and Analyze data >> generic keyword search

There is also a publication-related archive. To quote the site “Investigators are encouraged to deposit everything in the Publication-Related Archive that is needed to replicate the results of the study. This includes all data, computer programs, sets of computer program recodes, and extracts of existing data files used in the analysis along with the article itself. Investigators must include a description of the data and methodology that is detailed enough to permit another researcher to replicate the findings.” [3]

UK Data Archive

<http://www.data-archive.ac.uk/create-manage>

Similar in aim and scope to the Inter-University Consortia for Political and Social Research, the UK Data Archive is collection of data from the social sciences and humanities (largely funded by the ESRC, the JISC and the University of Essex).

There are top-level searches for

Major studies >> alphabetized

Latest data (new additions) >> filtered by released in the past x months

Advanced search >> keyword (limited by fields title/subject/keyword)

---Keywords

Geography
Data Collector/Funder
Methodology
Publications

Study ID

The site links out to other, international data archives, and as appropriate, by data set unit to related resources.

Archives of Teaching Materials:

Language Materials Library (Teaching resources)
<http://lmp.ucla.edu/>

Not all materials are data. The Inter-University Consortia for Political and Social Research offered documentation archiving. However, some libraries are designed with untraditional materials in mind. UCLA's Language Materials Library houses resources for language teachers that may be searched with the audience, level of difficulty and type of material in mind.

At the top-level search by (separate pull down menus to filter the criteria)

- Language
- Material (type of teaching resource)
- Level (beginner to advanced)
- Audience (all/post secondary/secondary/primary)

A side bar offers quick links to descriptions of content

- Teaching materials
- Authentic materials (intended for native speakers)
- Language profiles

And also a themed collection

- K-12 gateway

Summary

	Teatro alla Scala	LOC	Milman Parry	ITMA	Gospel Music History	Global Music Archive	Macaulay	RCD	ISMIR	IUPSR	UK Data Archive	LML
Restricted # of Formats							X	X	X			
Series; Themed Child-Collections		X		X						X		X
Study Unit level search	X	X	X	X	X	X	X					X
Variable Access Permissions or Access		X			X	X				X	X	
Variable Usage rules								X				
Strong curatorial message	X		X	X	X							
General and Academic Audience	?	X	?	X	X		X					
Academic Audience Only						X		X	X	X	X	X
Specialized Vocabulary				X			X			X	X	
Atypical Materials										X	X	X

Observations

The following early observations may be drawn from these use studies:

- Simplicity, ease of use and clarity are gained when a subject has a limited number of associated data types (Macaulay, audio and video only)
- Not all data sets require study unit searches/access (RWC database)
- Value to researchers of declaring the methodology (IUCPS, UK Data Archive)
- In small repositories, strong a curatorial message goes a long way towards increasing the value of the archive (USC Gospel Music History Archive)
- Features like the interactive score, simple, drastically increase the value of the archive and/by broadening the audience (ITMA)
- The importance of determining the level of expertise necessary to search. (What must the user know?)
- The importance of predict how the content will being used. Materials like handbooks and teaching materials, may work best w/dedicated search functions

- Collections of collections (or series) are unwieldy to browse. Researchers must know what they need.
- An academically-oriented archive may be given a broader appeal through the inclusion of series or themed children-collections.
- Making the search terms readily apparent (for example by using text fields) encourages non-expert to search by these criteria.

Next Steps

Restricted data is a major concern for AIRS. Several AIRS projects collect data from children including audio and video data. These projects are being conducted at different institutions and in different countries. Researchers are bound by different ethics committee guidelines and national and international laws relating to collecting data from human subjects and IP. Also, given that a lot of the audio and visual data collected through AIRS is musical in nature, IP may factor in in several ways. The Robertson Library's VREs do have built in functionality for collecting access information, tagging permissions and policy conformance.

Establishing a specialized vocabulary will be an essential part of the AIRS DL design process. The survey to researchers is the first step.

Refining the purpose and audience are essential components of the design strategy. Will AIRS DL offer resources for the general public? What do we expect the user to know before searching the system? What role does data set sharing have in the final design? Will sharing data take priority over the archiving of other resources?

References

[1] DDI: <http://DDIalliance.org> last retrieved 21 Feb 2012

[2] [Managing and Sharing Data: a Best Practice Guide for Researchers](#) (pdf).

Published by Data Archive UK. <http://www.data-archive.ac.uk/create-manage> last retrieved 22 Feb 2012

[3] Inter-University Consortia for Political and Social Research:

<http://www.icpsr.umich.edu/icpsrweb/content/ICPSR/access/deposit/guide/> last retrieved 22 Feb 2012