

A Casebook for Revitalizing Legacy Databases / Data Sets

This project will create a casebook for use in evaluating and migrating legacy databases and data sets present in libraries. The casebook will include: an introduction; several cases or examples of legacy databases and data sets needing migration that are currently in use in Florida libraries; and a recommendations and resources section.

Florida, and thus Florida libraries, has been a forerunner in many technological areas. The early adoption of new technologies resulted in implemented technologies that were quickly superseded. Lack of resources prevented the migration of many of these to newer technologies, and this is especially the case for small, legacy databases. These small databases often contain valuable information in the form of simple data sets; however, the databases in their current form are often both difficult to access at all and difficult to use. In many cases, the legacy databases could now be migrated to simple technologies that would better enable access and use, and that would have lower resource demands than are required to support the legacy systems in addition to other systems. These databases need to be migrated to ensure the content is in a sustainable format in terms of digital preservation and cost control. These databases also need to be migrated in order for their contents to meet ADA accessibility requirements for web access and to be generally findable, accessible, and useful in a modern, web-scale world.

While there is a clear need to migrate these databases, the process of migration is unclear. This casebook will include several databases, at least one of which that will be migrated by the end of the project. The casebook will include: a full case summary for each database to explain the context in which the database was created; its current need and reasons for use; current costs; obstacles to migration; migration costs; and resources for migration. The draft casebook will be shared with the State University Libraries' Digital Initiatives and Services Committee for review and use in planning statewide initiatives, services, and training relating to legacy databases and data sets. In creating the casebook: at least one legacy database will be migrated; at least two other legacy databases will be fully documented for future migration; the State University Libraries' Digital Initiatives and Services Committee will participate in reviewing and refining the casebook into a statewide resource; and the State's critical knowledge mass in handling legacy databases and data sets will increase.

Draft Casebook for Revitalizing Legacy Databases / Data Sets (2011)

Note

This historical project documentation is archived and available in case useful for historical purposes. Please note: the information is not current and was based on a brief project from 2011. For current data work, including legacy databases and datasets, see the UF Data Management/Curation Task Force: <http://library.ufl.edu/datamgmt>

As originally written, this is the initial draft of the *Casebook* was envisioned for review and use as a starting point in conversations with State University Libraries in Florida for shared needs in planning statewide initiatives, services, and training relating to legacy databases and data sets.

Abstract

The *Casebook* is intended for use in evaluating and migrating legacy databases and data sets present in libraries. The *Casebook* includes: an introduction section; cases or examples of legacy databases and data sets needing migration that are currently in use in Florida libraries; and a recommendations and resources section.

Florida, and thus Florida libraries, has been a forerunner in many technological areas. The early adoption of new technologies resulted in implemented technologies that were quickly superseded. Lack of resources prevented the migration of many of these to newer technologies, and this is especially the case for small, legacy databases. These small databases often contain valuable information in the form of simple data sets; however, the databases in their current form are often both difficult to access at all and difficult to use. In many cases, the legacy databases could now be migrated to simple technologies that would better enable access and use, and that would have lower resource demands than are required to support the legacy systems in addition to other systems. These databases need to be migrated to ensure the content is in a sustainable format in terms of digital preservation and cost control. These databases also need to be migrated in order for their contents to meet ADA accessibility requirements for web access and to be generally findable, accessible, and useful in a modern, web-scale world.

While there is a clear need to migrate these databases, the process of migration is unclear. The *Casebook* includes: a full case summary for each database to explain the context in which the database was created; its current need and reasons for use; current costs; obstacles to migration; migration costs; and resources for migration.

Casebook for Revitalizing Legacy Databases / Data Sets

Contents

- Introduction
- Organization and Content Overview
- How to Use the Casebook
- Cases
 - Case 1: Mickler-Goza Newspaper Article Database, 1762-1885
 - Case 2: Florida Newspaper Project Holdings Database
- Recommendations and Further Resources

Introduction

Florida, and thus Florida libraries, has been a forerunner in many technological areas. The early adoption of new technologies resulted in the implementation of technologies that were quickly superseded. Lack of resources prevented the modernization of many of these to newer technologies, and this is especially the case for small, legacy databases. In Information Technology a *legacy system* is a system that still has value, but that presents problems in its current form and is resistant to modification and evolution. Legacy databases thus contain valuable data, but the technical aspects of the databases have various problems at simple levels (e.g.; user interface is difficult to access and use) and in terms of the larger infrastructure (e.g.; database design) and neither can easily be corrected. The problem of legacy systems continues to grow unless the software is actively supported on an ongoing basis because, as explained by Lehman's Laws, "A large program that is used undergoes continuing change or becomes progressively less useful."¹ Many systems are designed to be evolvable, but that requires more time during the initial development. Legacy systems must be modernized into a form that is evolvable to prevent the same problems from recurring.

The State University Libraries in Florida maintain many legacy databases because the libraries function as both research units and research support units.² As such, the libraries undertake their own research projects as well as collaborative research projects with and in support of their faculty. Many of the libraries' legacy databases were developed for grant projects that have now ended and so funding is no longer available for ongoing maintenance or for modernization. Because the majority of these databases were developed specifically to support patron access to collections and related information resources, the data they contain remains important. These databases need modernization because their current technical implementations severely inhibit the information they contain from being accessed by patrons or even by other systems. They also need modernization to ensure their contents are in a sustainable format in terms of digital preservation and cost control (with duplicative costs for maintaining legacy and modern systems). Because many of the existing legacy databases developed in the early years of the Internet, they also fail to meet current standards for technical use and interoperability. Modernization is thus required to meet ADA accessibility requirements and to be generally findable, accessible, and useful in a modern, web-scale world.

In many cases, the legacy databases could be modernized through migration to new technologies. This migration would better enable access and use, and would have lower resource demands than are required to support the legacy systems in addition to other systems. While there is a clear need to migrate many of these databases, the process of determining which to migrate and how to do so is unclear. This *Casebook* was developed to aid in the modernization process.

¹ Lehman, M.M. & Belady, L. *Program Evolution: Processes of Software Change*. London: Academic Press, 1985.

² The problem of legacy databases is prevalent in all fields and in most libraries. According to a 2007 report, the Smithsonian Museum of Natural History had more than 22 legacy record systems in operation and in need of migration (http://fedtechmagazine.com/article.asp?item_id=273).

Organization and Content Overview

The *Casebook* provides background information and resources on data curation of information held in legacy databases along with specific case studies of legacy databases in use in State University Libraries in Florida. A casebook, as defined in WordNet, is “a book in which detailed written records of a case are kept and which are a source of information for subsequent work.”³ Casebooks are frequently used in law and medicine.⁴ All fields regularly use case studies because of the usefulness of documenting all aspects of contexts of a specific case that embodies core issues and can refer to specific resources.

Case studies are particularly important for digital curation issues faced by cultural heritage institutions. The recently published *Digital Forensics and Born-Digital Content in Cultural Heritage Collections* emphasizes the importance of case studies by including the collection of stories and case studies as one of the report’s eight recommendations.⁵ Another of the report’s recommendations is facilitating training. Casebooks answer both of these recommendations because they case studies they collect case studies are well-suited for use in teaching and training.

Because legacy databases implicitly require added complexity for their correction, they also face typical Information Technology modernization issues.⁶ Thus, each case study contains information related to database modernization strategy. Database modernization strategy includes flexibility, reuse, and interoperability as key technical goals.⁷ Each case study captures information on staff and affiliates with critical knowledge because database modernization strategy recognizes the importance of senior staff knowledge in the modernization process: “Senior workers represent a precious resource not only for understanding, maintaining, and integrating aging software systems, but also for replacing them with new technology that will enable the company to move forward without compromising current operations.”⁸ In Florida’s State University Libraries, many senior staff will soon retire, thus intensifying the need to begin modernization work or increasing the risk of needing to conduct more costly reverse engineering work.

Additionally, each case study includes a brief summary explaining: the context in which the database was created; current need and reasons for use; problems with the current database

³ <http://wordnetweb.princeton.edu/perl/webwn?s=casebook>

⁴ Casebooks have been used to teach law in the United States since the 1800s and their prevalence continues today with the recognition of their importance and usefulness underpinning several projects to develop open source digital casebook tools.

⁵ *CLIR pub 149: Digital Forensics and Born-Digital Content in Cultural Heritage Collections* by Matthew G. Kirschenbaum, Richard Ovenden, Gabriela Redwine with research assistance from Rachel Donahue; CLIR: December 2010 (<http://www.clir.org/pubs/abstract/pub149abst.html>). On author, Kirschenbaum, has noted the need for digital humanities and digital curation casebooks (c.f.; <http://otal.umd.edu/~mgk/blog/archives/000472.html>; www.samplerelity.com/2010/10/01/initial-thought-on-archiving-social-media/).

⁶ See: *Modernizing Legacy Systems: Software Technologies, Engineering Processes, and Business Practices* by Robert C. Seacord, Daniel Plakosh, and Grace A. Lewis; Boston, MA: Pearson Education, 2003.

⁷ <http://www.ibm.com/developerworks/webservices/library/ws-soa-legacymod/index.html>

⁸ <http://www.ibm.com/developerworks/rational/library/sep05/lieberman/>

structure and current costs in time and systems; obstacles to migration; migration costs; and resources for migration. For case studies that document a migrated database, the full migration process is also documented.

The second case study, the *Florida Newspaper Project Holdings Database*, also includes a summary of the database presented with the Data Curation Profile format recommended by Purdue in their Data Curation Profiles Toolkit.⁹ This format was selected for inclusion because the legacy databases in libraries are often developed from original scholarly research and are thus most closely aligned to typical scholarly data curation needs. A Data Curation Profile includes information about the data itself (lifecycle, purpose, forms, perceived value) and about the user needs for the data (accessibility, documentation required, preservation need). This information is useful in understanding the data in a sterile environment, without the added complexity required to extract, transform, and load the data into an evolvable system.¹⁰

The final section of the casebook contains recommendations and further resources. This section includes tips for how to evaluate and migrate legacy databases, a list of contacts in Florida for assistance in migrating legacy databases, and other resources gathered throughout the course of the project.

How to Use the Casebook

Each case study is an individual example of the problem, method, and solution for problems posed by a specific legacy database. The case studies are to be used to illustrate the full process of analysis, method, and final outcome. Each case study shows examples of requirements for migration, different possible solutions, true business costs and risks for maintaining legacy systems, and documentation needed through the evaluation and migration process. For new projects, this is helpful as a reference example to show requirements and workflow when seeking project approval of funding. In addition to the examples, the “References and Further Resources” section provides contacts and other resources.

As originally written, this is the initial draft of the *Casebook* was envisioned for review and use as a starting point in conversations with State University Libraries in Florida for shared needs in planning statewide initiatives, services, and training relating to legacy databases and data sets.

⁹ <http://www4.lib.purdue.edu/dcp/>

¹⁰ Extracting, transforming, and loading data requires data cleaning and data quality controls. See: Richard Wojcik, Logan Hauenstein, Carol Sniegowski, and Rekha Holtry’s “Obtaining the Data” in *Disease Surveillance: A Public Health Informatics Approach* edited by Joseph S. Lombardo and David L. Buckeridge; Hoboken, NJ: John Wiley & Sons, 2007.

Case 1: Mickler-Goza Newspaper Article Database, 1762-1885

Original: <http://www.uflib.ufl.edu/Goza/>

Migrated: <http://ufdc.ufl.edu/fdnlmg>

Context for the Database Creation

The *Mickler-Goza Newspaper Article Database, 1762-1885* (<http://www.uflib.ufl.edu/Goza/>) was created in 1999-2000 to provide an easy way for patrons to find news articles about Florida from the years before Florida had its own newspapers held in the University of Florida Collections.¹¹ The homepage for the database explains:

With the exception of the East-Florida Gazette in the 1780s and a small press at Fernandina in 1817, Florida had no colonial newspapers. Even in the immediate aftermath of cession in 1821, only a few newspapers served Florida. The Newspaper Article Database consists of stories and reports about Florida gathered together by the Goza and Mickler families and donated to the P.K. Yonge Library of Florida History. There are approximately 1500 articles in the database. They are all from non-Florida newspapers and cover events in Florida between 1762 and 1885. The articles pre-dating the Territorial Period help to "fill in" the journalistic record at a time when there was no Florida press, while the articles from after 1821 both complement and supplement news published in Florida.

The database itself was designed as a simple search engine containing:

- Date of the newspaper issue
- Name of the newspaper in which the article appeared
- Page and column number(s) of the article
- Description of the article and/or title
- Link to full transcribed text, when available

The database includes 1,530 entries and links to 214 full text articles.

¹¹ See the "Find out more about this database" page: <http://web.uflib.ufl.edu/spec/pkyonge/micgoz.html>

Mickler-Goza Newspaper Article Database, 1762-1885: Homepage

University of Florida
George A. Smathers Libraries

Hours | Ask a Librarian | Online Requests | Remote Logon
Library Catalog | Databases | Site Map | Help | Search

[Library](#) > [Special & Area Studies Collections](#) > [PK Yonge Library of Florida History](#)

Newspaper Article Database, 1762-1885

From the Collections of
William and Sue Goza and Thomas and Georgine Mickler

[Find out about this database](#)

Enter the SEARCH values for the fields you wish to search.

Date	From: <input type="text"/>	To: <input type="text"/>
Newspaper Name	<input type="text"/>	
Description	<input type="text"/>	

[Return to the P.K. Yonge Library of Florida History](#)

Mickler-Goza Newspaper Article Database, 1762-1885: All Items, from Empty Search

University of Florida
George A. Smathers Libraries

Hours | Ask a Librarian | Online Requests | Remote Logon
Library Catalog | Databases | Site Map | Help | Search

[Library](#) > [Special & Area Studies Collections](#) > [PK Yonge Library of Florida History](#)

Newspaper Article Search Results

(in ascending order by date)

There are 1530 records that match your criteria.

Date	Newspaper Name	Pages	Description	Newspaper Link
2/2/1762	London Chronicle	107:1	Oglethorpe's seige (retrospective). "When Oglethorpe led the British troops against Augustine, the terror of our Southern settlers . . ." Mickler.	N/A
8/8/1768	Boston Chronicle	314:3	Extract of a Letter from Pensacola, June 10, Concerning settlement of Virginians in West Florida. Goza.	Image
11/7/1777	Lloyd's Evening Post	442:1-2	Extract of a Letter from St. Augustine. Account of a skirmish in Florida on Nassau River during the American Revolution. Mickler	N/A
4/18/1781	New Jersey Gazette	1:3	Galvez's siege of Pensacola. "On the morning of the 9th of March, a Spanish fleet appeared off the bar of Pensacola." Mickler	N/A

Current Need and Reasons for Use

The database is needed to provide access to the index of these historic news accounts of Florida and the links to full text articles where applicable. For the articles without full text, some are available on microfilm reels that can be borrowed through interlibrary loan. Others are only available in their original print form and patrons must travel to the University of Florida Libraries for access. Thus, the database allows patrons to find historic news accounts of Florida and access the full text in some instances.

Problems with Current Database Structure and Current Costs

The current database shows its age. It is not designed for the current web-scale world driven by search engine access and linked information. It has not been updated to a more modern design and the initially limited data structure has been unchanged. While the data for each newspaper article is separated into multiple fields, the data structure is very limited. First and foremost, the information in this database is divorced from other relevant information, such as information on which articles are available on microfilm. To see if particular articles are available on microfilm or if any additional items have been digitized, researchers must either: search the UF Libraries' catalog or to contact someone in the UF Libraries. Requiring the patron to search the catalog, particularly with only brief information to guide the search, incurs a quality of service cost with a negative impact on patron services. Requiring the patron to contact someone in the UF Libraries is likely to result in excellent customer service with the patron learning a great deal about the resources of interest and related materials. However, this personalized service incurs a cost in terms of staff resources for the UF Libraries and it is a cost better served through an improved database where patrons would have ready access to answers without needing to personally inquire.

The data contained in the database is also of limited use. The results cannot be sorted after display with ascending order by date as the only display option. Sorting for ascending and descending order for all data columns is a normal database functionality expected by most users.

The explicit system costs for supporting the database are minimal. The UF Libraries run a Microsoft SQL Database server which runs a number of small databases. The *Mickler-Goza* database is on this shared server and so the costs to operate the *Mickler-Goza* database are subsumed into the costs of operating the other databases. It is not an additive cost.

While the explicit costs are minimal, the true costs for the *Mickler-Goza* database include factors present with many legacy databases:

- Negative impact or cost to the quality of service;
- Lost opportunity costs; minimal return on investment (ROI) for creating and maintaining the database; and,
- Potential cost with the risk of loss, which is always a concern with data that is not being actively maintained.

This database was not planned to be included in the *Casebook*. However, it came to be included as a matter of necessity in February 2011, during the initial timeline for creating the first draft of the *Casebook*, after a patron alerted the UF Libraries to a problem with this database. Using the UF Libraries' web contact form for Special Collections, a patron reported:

I have been trying to access the online images for the Historical Newspaper database for about one week. At first, I received a message that my browser could not open the page. I downloaded a different browser, which was also [message clipped in original]

While the patron had noted that the inquiry was intended for Special Collections, the web contact form is processed by the Information Technology Department. They routed the patron inquiry to the UF Digital Collections. The UF Digital Collections include several newspaper digital collections, all of which were operating properly. The UF Digital Collections often receive inquiries unrelated to the UF Digital Collections and so responded to the patron with a request for additional information, specifically requesting a link to the error page if possible for additional testing and support. The patron responded with a link to the *Mickler-Goza Newspaper Article Database, 1762-1885* (<http://www.uflib.ufl.edu/Goza/>). The UF Digital Collections did not have experience or knowledge of this database. The work to correct the problem first required research before a corrective strategy or solution could be developed.

In researching the problem, the citation database component was found to be working properly, albeit in the same limited manner in which it originally operated in 2000. However, the links to full text were not working. The links went to resources hosted at the statewide Florida Center for Library Automation (FCLA). Inquiring with FCLA, FCLA explained that the computer server that was hosting the files had been decommissioned. The content on the server that was planned for migration had been migrated to new systems. FCLA had no indication that the text files for the *Mickler-Goza Newspaper Article Database, 1762-1885* were to be migrated. Instead, the files were in the process of being deleted.

Luckily, the deletion process was incomplete. FCLA was able to provide a single SGML file containing the full text for all files pending deletion. The SGML file contained the full text for all of the articles from this legacy database, as well as the full text for a number of other materials. The SGML file was a single, massive file.

Without the SGML file, the true costs of this legacy database would have increased by the costs to re-transcribe 214 articles. The true costs had already increased in terms of negative impact to patrons, and staff time for patron support and the immediate problem response. Perhaps most telling of the risks involved with legacy databases is the fact that the Libraries would not have known about the problem if the patron had not reported it. By existing alone and operating outside of other supported systems, the legacy database had fallen into neglect. The neglect failed to support any benefit from the work already done and allowed the possibility of data loss to occur.

Obstacles to Migration

The database appears to have been developed as a lower-cost alternative to cataloging the newspaper issues and articles. It appears as though the database and full text file hosting was

implemented as a temporary and incomplete solution using the most affordable technical option as a stop gap until it could be replaced as part of a larger project. As a temporary and incomplete solution, normal supports were not in place. The database was not known to the faculty and staff supporting the UF Digital Collections, who are the primary contacts for files hosted by FCLA. It appears that the database's placement outside of any normal support channels and staff changeovers at both institutions allowed communication to cease entirely regarding these materials, resulting in the perception that the files were not needed and could be deleted. Initial resource scarcity impact on database design, operation outside of normal channels, and staff created a situation where neglect was able to occur. These factors resulted in the near deletion of the text files and made the migration more difficult by requiring a large amount of research time for correction.

Migration Costs

In addition to the less quantifiable costs in terms of service impact and risk of data loss, the migration process required approximately 100 hours of highly skilled faculty and staff labor.

The system costs were negligible because the UF Digital Collections system could provide all needed functionality for supporting the full text files. The UF Digital Collections' operating costs are a highly stable cost and the small number and size of these files represented a negligible cost.

Resources for Migration and Database Modernization Strategy

Database modernization strategy focuses on flexibility, reuse, and interoperability. The primary resource for migration following the principles for database modernization strategy was the availability of the UF Digital Collections for full system support. The *Mickler-Goza* legacy database was well positioned within the UF Libraries for support because of so many similar projects supported on the UF Digital Collections. The UF Digital Collections was not just an appropriate system; it was an optimal system for full support.

The *Mickler-Goza* database was created before the UF Digital Collections began in 2006. The UF Digital Collections was specifically designed to support the needs of libraries, archives, and museums, including the integration and expansion of existing standalone systems as appropriate. Because of the excellent technologies and functionalities available through the UF Digital Collections, all systems appropriate for integration are migrated into the UF Digital Collections as resources allow. The quality of the UF Digital Collections as a system also brings a large internal and patron user community, ensuring institutional and patron-drive support for the ongoing care of materials.

Migration Process

The migration process included:

- Responding to the patron
- Researching the database and the full text files
- Obtaining the SGML file

- Parsing the SGML file to remove extraneous files
- Parsing the 214 articles within the single SGML file into the 214 appropriate individual files
- Normalizing the files for ingest into the UF Digital Collections
- Ingesting the files into the UF Digital Collections
- Updating the links in the database to link to the new, permanent file locations
- Cataloging the 214 full text articles that were migrated and a portion of the others
- Planning for support to complete the catalog records for the remaining articles

To process the files into the UF Digital Collections, all of the individual files had to be parsed (separated into the individual files) from the single SGML file. Then, the individual files had to be converted into standard formats so that they could be added to the processing queue for processing, loading, and archiving.

The technical processing began with evaluating the full SGML and all of the included files. Reviewing the file confirmed that only the *Mickler-Goza* newspaper files required migration. After parsing the SGML file for only the 214 needed files, the SGML file was parsed again into the 214 discrete text files. The 214 files were converted into PDF and other normalized derivative files to support optimal online access and long-term preservation. The files were linked to basic metadata records using the available information. Then, the files as bundled with the metadata were ingested into the processing queue for loading to the UF Digital Collections. This process would normally have included two stages: one to create the framework for conversion and processing, and another to have student workers conduct the processing. Separating the work into two stages is also a slower process. The two stages require more time overall with less time from faculty and staff which results in a lower overall labor cost. In this particular instance, there was not a labor saving method for separating the work into two stages. If separating the work into two stages had been possible, the cost savings would have been weighed against the value of timely correction given that the database had failed and patrons were being impacted by a reduced service level.

After the files loaded into the UF Digital Collections, the Information Technology and Cataloging Departments were contacted for follow-through on the remaining work requirements. The Information Technology Department was given the new permanent links to update the links in the database and they were able to do this process quickly. The Cataloging Department was contacted regarding the possibility of enhancing the existing records to support findability and usability. Cataloging requested and received additional short-term funds for hourly staffing to update the records.

Case 2: The Florida Newspaper Project Holdings Database

<http://www.uflib.ufl.edu/fnp/>

Florida Newspaper Project Holdings Database: Homepage



**UNIVERSITY OF
FLORIDA**
GEORGE A. SMATHERS LIBRARIES

The Florida Newspaper Project - Search Form

Holding Institution:

Publication Name:

Publication City: County:

Sort results by:

☒ Publication Location ☐ Publication Name ☐ Holding Institution

All searches are performed by **pattern-matching** in the indicated fields. Boolean searches are not supported.



This database was compiled as part of the [Florida project in the United States Newspaper Program](#),
funded by the [National Endowment for the Humanities](#).

Florida Newspaper Project Holdings Database: All Items, from Empty Search



**UNIVERSITY OF
FLORIDA**
GEORGE A. SMATHERS LIBRARIES

The Florida Newspaper Project - Linked Search Results

Holding Institution:	Publication:	Holdings:	Pub. City:	Pub. County:
Orange County Museum	AAFSATONIAN	has issues: 1943: Apr. 10,24; May 1-29; June 6,19-26; July 3-10,31; Aug 14.	1	2
State Library of Florida	AAFSATONIAN	has microfilm and originals: 1943: Sept.11-Oct 30.	1	2
Florida State University (FSU)	Alachua Post	has microfilm: 1907: Jan. 11(Supplement).	Alachua	Alachua
State Library	Alachua Post	has microfilm: 1907: Jan. 11(Supplement).	Alachua	Alachua
University of Florida (UF)	Alachua Post	has microfilm: 1907: Jan. 11(Supplement).	Alachua	Alachua
University of West Florida (UWF)	Alachua Post	has microfilm: 1907: Jan. 11(Supplement).	Alachua	Alachua
University of West Florida (UWF)	Alachua Times	has issue: 1941: Nov 13.	Alachua	Alachua

Florida Newspaper Project Holdings Database: Individual Record Example for Alachua Post

University of Florida	Hours Ask a Librarian Online Requests Remote Logon
George A. Smathers Libraries	Library Catalog Databases Site Map Help Search

The Florida Newspaper Project - Publication Details


Publication:	Alachua Post
LC Control #:	sn95-47165
Title Variation:	-
City:	Alachua
County:	Alachua
Publisher:	Unknown
Current Frequency:	Unknown
Former Frequency:	-
Began Publication:	1???
End of Publication:	19??
Notes:	-
Citation:	-
Also Available:	-
Language:	-
Topics:	-
See also:	-
Predecessor:	-
Successor:	-

Staff Web | Staff Directory | Departments | Privacy Policy

Send suggestions and comments to: lib-webmaster@uflib.ufl.edu
© 2004 - 2006 University of Florida George A. Smathers Libraries.
All rights reserved.
Acceptable Use, Copyright, and Disclaimer Statement
Last updated February 15, 2004 - tlm

UF UNIVERSITY of FLORIDA
The Foundation for The Gator Nation

Data Curation Profile – Information Sciences/Digital Libraries

Profile Author	Laurie Taylor
Author's Institution	University of Florida
Contact	Laurien@ufl.edu
Researcher(s) Interviewed	Winston Harris
Researcher's Institution	University of Florida
Date of Creation	January 17, 2011
Date of Last Update	July 6, 2011
Version of the Tool	1.0
Discipline / Sub-Discipline	Information Sciences / Digital Libraries
Sources of Information	An initial interview conducted January 17, 2011 Review of existing database; full sample of the profiled data
Notes	Part of a larger study on legacy database modernization needs
URL	Forthcoming
Licensing	Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License. 

Section 1: Brief summary of data curation needs

The data set consists of a simple SQL database which is hosted by the centralized CNS/Open Systems group at UF (<http://open-systems.ufl.edu/hosting>). The data is currently not crawled by commercial search engines and many people do not know how to access it. For those that do, the data can be confusing because it is such a small data set and is presented in absence of the digitized holdings and other catalog records.

Section 2 - Overview of the research

The Florida Newspaper Project Holdings Database developed from the Florida instance of the US Newspaper Project by the Library of Congress and the National Endowment for the Humanities. Participants in the US Newspaper Project located and microfilmed state newspapers held by many different institutions to ensure complete runs whenever possible. The University of Florida was the lead institution for the State of Florida and developed the Florida Newspaper Project Holdings Database as part of the project.

Because of changes to the cataloging for microfilm reels and because of the complexity of newspaper serial records, the information in this database is still useful and needed for locating specific reels and for planning digitization projects. There is a need to preserve and provide access to this data.

2.1 - Research area focus

The research goals are locating newspapers held by various institutions and then using this information to compile and preserve those information resources.

2.2 - Intended audiences

Libraries, archives, museums, historical societies, newspaper publishers, researchers, and the public interested in resources on Florida; especially cultural heritage institutions for preservation and access.

2.3 - Funding sources

Library of Congress; National Endowment for the Humanities; University of Florida Libraries

Section 3 - Data kinds and stages

3.1 - Data narrative

All data is finalized at this time.

3.2 – The data table

- Data Stage:

All data is finalized at this time. “Finalized” is the last stage in the data lifecycle in which all re-workings and manipulations of the data by the researcher have ceased.

Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
Primary Data				
Raw	n/a	0	n/a	
Processed	n/a	0	n/a	
Analyzed	n/a	n/a	n/a	
Finalized	all	1.39MB; 2,358 records	SQL db	
Ancillary Data				
Ancillary Data	n/a	n/a	n/a	

3.3. - Target data for sharing

All data should be shared in an accessible manner that is conducive to discoverability online.

3.4 - Value of the data

Because of changes to the cataloging for microfilm reels and because of the complexity of newspaper serial records, the information in this database is still useful and needed for locating specific microfilm reels and for planning digitization projects.

3.5 - Contextual narrative

The data is not up to date. If the data was discoverable through a general web search, it would be more used as-is and more frequently updated.

Section 4 - Intellectual property context and information

4.1 - Data owner(s)

This data is in the public domain.

4.2 - Stakeholders

Contributors, funding agencies, researchers, and the public.

4.3 - Terms of use (conditions for access and (re)use)

This data is in the public domain and can be used without restriction.

4.4 - Attribution

This data is in the public domain and can be used without restriction. Attribution with a link to the new online data location would be desirable to better connect others with this data.

Section 5 - Organization and description of data (incl. metadata)

5.1 - Overview of data organization and description (metadata)

The data is currently openly available online. The data for each record includes partial bibliographic and holding records.

5.2 - Formal standards used

The data is in a SQL database and some was derived from MARC records.

5.3 - Locally developed standards

N/A

5.4 - Crosswalks

N/A

5.5 - Documentation of data organization/description

None.

Section 6 - Ingest / Transfer

N/A

Section 7 – Sharing & Access

7.1 - Willingness / Motivations to share

Yes. This is intended.

7.2 - Embargo

None.

7.3 - Access control

None.

7.4 Secondary (Mirror) site

Not needed if data preserved and supported with general operational time norms.

Section 8 - Discovery

The database currently allows anyone to search the database online and to search for newspapers by: publication location; publication name; and holding institution. Open online access is required functionality; ideal functionality would be full text access for users and web robots.

Section 9 - Tools

N/A

Section 10 – Linking / Interoperability

Applicable for secondary stage where data would be reintegrated with digitized items.

Section 11 - Measuring Impact

11.1 - Usage statistics & other identified metrics

Usage statistics desired at level of normal web statistics.

11.2 - Gathering information about users

Not of interest; would conflict with privacy protections for patrons.

Section 12 – Data Management

12.1 - Security / Back-ups

Current and desired security and back-up practices must follow campus IT policies for data of value.

12.2 - Secondary storage sites

Only needed as part of redundancy for preservation.

12.3 - Version control

Not needed at this time; may become applicable after data is accessible and being utilized.

Section 13 - Preservation

13.1 - Duration of preservation

The data needs to be retained permanently. If integrated with other permanent data repositories/systems, this data version could be removed.

13.2 - Data provenance

Not necessary for existing data; would be needed for updates.

13.3 - Data audits

N/A

13.4 - Format migration

The data needs to be accessible on a web scale and needs to be integrated with other systems, not in its own silo, and it cannot incur new costs. The explicit current system costs for supporting the database are minimal. The UF Libraries run a Microsoft SQL Database server which runs a number of small databases. The *Florida Newspaper Project* database is on this shared server. The operational costs are subsumed into the costs of operating the other databases. It is not an additive cost.

For integrating this information with an existing system, the best system option is the UF Digital Collections. Migration costs entail adding a data view to the material views for items in the UF Digital Collections. This entailed time donated from a programmer from Boxing Clever to create the programming for a data viewer which was donated to the UF Digital Collections. Implementation of this programming is a nontrivial task and will require 20 hours of time from the UF Digital Collections programmer. This work is pending time availability for implementation.

Section 14 – Personnel

The data is maintained as part of the SQL database and the database administrator is the only support person.

14.1 - Primary data contact (data author or designate)

Winston Harris, UF Libraries

14.2 - Data steward (ex. library / archive personnel)

Winston Harris, UF Libraries

14.3 - Campus IT contact

Winston Harris, UF Libraries

14.4 - Other contacts

Winston Harris, UF Libraries

Recommendations and Further Resources

Contacts for assistance in migrating legacy databases:

- Digital Library of the Caribbean (dLOC): <http://dloc.com/contacts>
- Dataset support in SobekCM: <http://ufdc.ufl.edu/AA00017907/00001/pdf>
 - SobekCM: <http://sobek.ufl.edu/>
 - SobekCM Google Group: sobekcm-discuss@googlegroups.com

Additional Resources:

- Paradigm on digital private papers for evaluating digital resources and records: <http://www.paradigm.ac.uk/workbook/>