Data Curation Profile – Linguistics / Etymology

Profile Author(s)	Sonia Wade Lorenz & Lisa Zilinski
Author's Institution	University of South Florida in Lakeland
Contact	Lisa Zilinski, lzilinski@usf.edu
Researcher(s) Interviewed	Professor of English Composition & Literature [name withheld] Professor of Psychology [name withheld]
Researcher's institution	University of South Florida in Lakeland
Date of Creation	December 30, 2011
Date of Last Update	June 6, 2012
Version of the Tool	Data Curation Profile Toolkit is 1.0
Version of the Content	Profile Content: 1.5
Discipline/Sub-Discipline	Linguistics / Etymology
Sources of Information used for the profile	• An initial interview conducted on December 22, 2011. • A worksheet completed by the scientists as a part of the interviews. • A recording of the interview on December 22, 2011.
Notes	The profile was completed prior to the start of the researchers' experiment. The experiment is not scheduled to begin until the Fall semester of 2012. Since the researchers are still in the planning stages of the research, the scope of the profile may be modified significantly as the research progresses. Any modifications of this document will be subject to version control, and annotations require a minimum of creator name, data, and identification of related source documents.
URL	http://datacurationprofiles.org
Licensing	This work is licensed under a Creative Commons Attribution 3.0 Unported License

Brief summary of data curation needs

The researchers will be generating a significant amount of data but lack the means of managing, curating and sharing this data with others as effectively as they would like. The data will be composed of videos, spreadsheets and a finalized report captured in MS Word files. The project has not yet started so there is no data at this time. The researchers are very interested in making the video content/data available to others from the beginning, but they prefer to keep their interpretation of the data (Excel/SPSS spreadsheets and finalized reports) private for five years after which only the University of South Florida in Lakeland (USFL) will have access to them.

The researchers are applying for additional funding to address issues with the data. This funding would also help to create a digital asset management system for the university as a whole.

http://dx.doi.org/10.5703/1288284315008

Overview of the research

Research area focus

The research project is being co-developed by the above listed Psychology and English professors and is designed to determine the evolving meaning of the term “polytechnic”. USFL will be accepting their first freshmen class in Fall 2012. As a standard class assignment, freshman students will be required to execute video interviews of upper classmen to find out what they think “polytechnic” means. This assignment will be repeated with each group of incoming freshmen so that all students will have the opportunity to interview and be interviewed. By interviewing the students each year, the researchers will be able to understand the changing definition of the term “polytechnic” as the students progress through school. By comparing responses from differing freshmen classes, the researchers will be able to understand the changing definition of the term “polytechnic” as it is understood by the general public.

As the videos are entered into the repository, metadata will be entered to aid in determining common themes. Common themes will be extracted and logged into Excel spreadsheets. The results logged in the spreadsheets will be interpreted and reported in formal reports at the preset review times. The data stored in the spreadsheets will be reviewed and compared in the formal report after one year, five years and ten years.

Intended audiences

The researchers indicate that other polytechnic campuses, policy-makers, and higher education schools are the intended audience.

Funding sources

The primary funding agency for this project is USFL. However, the researchers are applying for a grant through the National Science Foundation (NSF). The researchers have indicated that the grant will require them to store the initial data/content (the videos) in a repository, but they do not believe they are required to store their interpretation of the raw data/content (Excel sheets and reports) in a repository. At this time, they do not believe a data management plan is required.

Data kinds and stages

Data narrative

The study is designed to review the evolving definition of the term “polytechnic” over a 10 year period. During the 10 year study, each new class of first year students will be required to create video interviews of other students (all class levels). The person being interviewed must describe what they believe the term “polytechnic” means. The video interview is the raw data/content.

The videos will be entered into a digital repository with appropriate metadata. The metadata will highlight the primary theme of the interview. Metadata will be entered by either the student and/or the supervising professor. After the videos with metadata are uploaded into the digital repository, the themes will be entered into an Excel spreadsheet by a graduate assistant. After the first year, fifth year and tenth year, the data listed in the Excel spreadsheets will be examined. The researchers will then report on their findings in a research report. Each report will build on previous findings.

The data table

Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
Primary Data				
Raw	Student Interviews	150 MB Each	Video	We expect to generate 100–150 files per semester
Processed	Video and Meta data entered into digital repository	150 MB Each	Video with Metadata	We expect to generate 100–150 files per semester
Analyzed	Themes extracted from repository and placed in spreadsheets	10 MB	Excel & SPSS	
Finalized	Report	30-50 KB	Word	

Note: The data specifically designated by the researchers to make publicly available are indicated by the rows shaded in gray. Empty cells represent cases in which information was not collected or the researchers could not provide a response.

Target data for sharing

The raw data (video content) will be shared openly in the digital repository and will be available to anyone in the processed format with metadata. The analyzed data (statistical sheets) and finalized reports would only be available to other researchers within the USFL after the fifth year report is completed.

Value of the data

The researchers indicate that other polytechnic campuses, policy-makers, and higher education schools are the intended audience. They may use this information to make policies and map education models.

Contextual narrative

The researchers indicate that the project is still in the planning stages and specific information regarding file size and the nature of the SPSS statistical spreadsheets are unavailable at this time. A second interview scheduled after the project begins will specify these details.

Second, in an effort to eliminate unnecessary variables and keep the project pure, more privacy is required in the first five years than in later years.

Finally, the researchers will initially house the video files on two hard drives in two separate physical locations until they can be entered in the digital repository.

Intellectual property context and information

Data owner(s)

The researchers advised that issue of intellectual property and ownership of the data has not been fully addressed. However, they do believe that they both (Lennon and Kelling) have primary ownership in the data.

Stakeholders

USFL may be considered a stakeholder in the data as USFL is the employing institution. Since the project is still in the planning stages, potential stakeholders, including graduate students and funding agencies, have not yet been determined but will be addressed in the second interview.

Finally, while students will be creating the video files, they will do so as part of their class assignment with the intent of posting the videos in the university's digital repository. They are not stakeholders in the data as a whole.

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

The researchers are still in the planning stages of their study and specific terms of use and reuse were not discussed.

Attribution

The researchers consider it a high priority that they be acknowledged if others were to use this data set.

Organization and description of data

Overview of data organization and description (metadata)

The researchers indicate they intend to enter the videos into a digital repository with metadata attached. The researcher's primary goal with the metadata is to create an easy way to identify and list themes within the digital repository so that the themes can be easily extracted and listed in the SPSS spreadsheet. At this time, controlled vocabulary for the metadata has not been determined. The researchers indicated that the video responses will direct the controlled vocabulary and metadata.

At this time, the researchers are not certain who will be entering the videos and metadata into the digital repository or who will be listing the themes in the SPSS spreadsheets. They anticipate this work will be handled by them, graduate assistants and/or the students.

Formal standards used

Not yet determined.

Locally developed standards

Not yet determined

Crosswalks

Not discussed.

Documentation of data organization/description

At this time, there are no existing documentation and/or documentation practices in place regarding the description and/or organization of the data.

Ingest / Transfer

The researchers indicated that students will need to be able to upload their videos into the digital repository and that everyone (internal and public) should have access to view the digital repository. The refined/analyzed data (SPSS spreadsheet) and finalized reports must only be available to immediate collaborators in the first five years. After the finalized report is completed in the fifth year, the analyzed data and finalized reports should only be available to researchers at USFL. The repository must support video files and SPSS.

Sharing and access

Willingness / Motivations to share

The researchers are willing to share the raw and processed data (videos) with anyone at any time, but they do have reservations about sharing the analyzed data (Excel and SPSS spreadsheets) and the finalized reports. They are particularly concerned with sharing the analyzed data and finalized reports in the early stages of the study, i.e.: the first five years. During the first five years, they only wish to share the analyzed data and finalized reports with their immediate collaborators. After the first five years, they are only willing to share the analyzed data and the finalized reports with researchers within USFL.

Embargo

The researchers do not require an embargo on their raw data. The analyzed data and finalized reports would require a 5.5 year embargo that would be lifted after the fifth year finalized report was completed.

Access control

The raw data will be publicly available at all times. The researchers will restrict access to the analyzed data and finalized reports for the first five years so that only immediate collaborators are able to review it. After the first five years, only researchers within the university will have access to the analyzed data and finalized reports.

Secondary (Mirror) site

Having a secondary mirror site that can be accessed if the repository is off line is not a high priority for the researchers. Having a backup storage site for the repository housed at a separate geographic location is a high priority.

Discovery

The researchers indicated that the ability for others to be able to search for the raw data (the videos) on Google is a high priority. Therefore, metadata is also a high priority. They do not want their analyzed data or finalized reports to be open to the public through Google or any other public access site. The analyzed data and finalized reports should be searchable within the university after the 5.5 year embargo is lifted.

Tools

To generate the data, students would require the use of video cameras, video editing software, and animation software (all equipment provided in USFL's DMIS lab).

To house and refine the data, the researchers would require external hard drives, a digital repository, SPSS, Excel, and Word.

Linking / Interoperability

Linking and Interoperability is a low priority for the researchers in general. However, they would like to be able to compare the internal analyzed data sets from multiple years within this study.

Measuring impact

Usage statistics and other identified metrics

The researchers stated that usage statistics are not a priority.

Gathering information about users

The researchers stated that information about users is not a priority.

Data management

The research project is not scheduled to start until Fall 2012; therefore, at this time there is no data. The researchers intend to store the data in a minimum of two hard drives (one on site and the other off site) and in a data repository.

Security / Back-ups

In the past, the researchers have stored information and data on their computers and then backed that information up using tools like an external hard drive and/or internet features like Drop-box.

For this project, they would like to store their raw data and processed data (the video content) on a minimum of two hard drives in separate locations and in the university's digital repository. They intend to use similar methods for storing the analyzed data and finalized reports; however, the analyzed data and finalized reports would be kept more secure with a limited audience whereas the video files would be open to the public.

Secondary storage sites

Having a secondary storage site at a different geographic location for the data is rated as a high priority by the researchers.

Version control

Having the ability to enable version control for the dataset is rated as a high priority by the researchers.

Preservation

Duration of preservation

The researchers indicated that their data should be preserved indefinitely. The data would be useful to compare to future responses for continued analysis.

Data provenance

Documentation of any and all changes made to the data over time is a high priority for the researchers.

Data audits

The ability to audit the dataset is a medium priority for the researcher.

Format migration

The ability to migrate the dataset into new formats over time is a medium priority for the researcher.

Personnel

Primary data contact

Professor of Psychology
University of South Florida in Lakeland

Professor of English Composition & Literature
University of South Florida in Lakeland

Data stewards

Lisa Zilinski
Business Librarian
University of South Florida in Lakeland

Sonia Wade Lorenz
Library Assistant
University of South Florida in Lakeland

Other contacts

N/A