Data Curation Profile –Public Health

Profile Author	C.Hudson
Author's Institution	Washington University in St Louis
Contact	C.Hudson chudson@wustl.edu
Date of Creation	April 3, 2012
Date of Last Update	
Version of the Tool	1.0
Version of the Content	
Discipline / Sub-Discipline	Public Health Communication
Sources of Information	• An initial interview conducted on April 2, 2012 by Sylvia Toombs, Social Work Librarian.
Notes	
URL	
Licensing	

Brief summary of data curation needs

The scientist is possibly interested in storing his data if some mechanism/requirement existed that would allow him to collaborate and develop new analyses of his data.

Overview of the research

Research area focus

The researchers is a leading national public health expert in the field of health communications. The research that is connected to this data has been collected for two years and is still ongoing. The research involves individuals who have contacted the United Way 211 information referral service to connect with community services, including assistance with finding a food pantry, housing, pro-active health screenings, etc. He collects data regarding individual life situations, health risks, and follow through on health and services recommendations.

Intended audiences

The scientist feels that other researchers, 211 systems, students, and other individuals with instruction opportunity would have an interest in the data.

Funding sources

Funding for this project came from the National Cancer Institute. He is required to have a data management plan that outlines data safety and monitoring, as well as accrual information.

Data kinds and stages

Data narrative

The scientist collects personal information data from individuals who use the United Way 211 referral service. Data is still being collected.

The raw data are collected in four different waves. The first wave of data is collected by 211 operators. Data collected at this wave includes a cancer risk assessment and the reasons an individual called 211. The second wave of data collection is also collected by the 211 operators. This data includes a baseline assessment of an individual's life situation. The third wave of data is collected by his employees at the one month mark. In this follow up, cancer risks are reassessed and problem resolution is discussed. The fourth wave of data is collected by public helath employees at the four month point. In this final follow-up, data regarding behavioral changes, 211 satisfaction and a number of other parameters are collected.

The second stage of the scientist's data lifecycle would be compiling all these raw data sources into one master database.

In the third stage the data manager cleans the data and checks for data quality and completeness. This happens every month to create the most up to date final data set.

The fourth data stage would be best described as an interim complete data stage. At this stage the data analyst applies a set of analyses and generates 16 different variable tables.

The fifth stage of the data would be the final stage. In this stage the study is over and completed and no new data is being added.

The data table

Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
Primary Data				
Raw	Data from various waves	1 / 16MB	Varies	
Compiled data	Compiled data of waves	1 / 16 MB	Varies	
Cleaned data	Cleaned and quality checked data	1 / 16 MB	Varies	
Interim complete data	Analyzed data	1 / 16 MB	Varies	
Final complete data	Data no longer being added	1 / 16 MB	Varies	

Note: The data specifically designated by the scientist to make publicly available are indicated by the rows shaded in gray (the "processed" row is shaded here as an example). Empty cells represent cases in which information was not collected or the scientist could not provide a response.

Target data for sharing

The scientist would be willing to share the data with immediate collaborators at the interim complete stage. He would be willing to share the final complete data incrementally over time upon completion of any additional analyses that immediate collaborators, others in the research institution, and others in his field have completed any analyses they wanted to research using his data.

Value of the data

The scientist sees educational and instructional value in the data. The datasets could be mined by undergraduates/graduates working on related thesis/dissertations. He also felt that the data would be useful for other 211 programs and anyone interested in 211 caller information.

Contextual narrative

The researcher indicated a willingness to share the data with others incrementally, but he was also interested in developing a policy that would have those individuals who used his data to share their research questions with the team, acknowledge or cite the original project, and grant the scientist a right of refusal on authorship.

Intellectual property context and information

Data owner(s)

The scientist indicated that he and his project partners are the owners of the data.

Stakeholders

He feels as though the United Way is also a stakeholder in the data.

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

As stated above, the scientist was interested in developing a policy that would have those individuals who used his data to share their research questions with the team, acknowledge or cite the original project, and grant the scientist a right of refusal on authorship.

The ability to create a basic, public description of his data was a high priority for the researcher.

The researcher felt it was a medium priority to restrict access to the data set to authorized individuals.

Attribution

The scientist felt it was a low priority to have the ability to cite the dataset in his publications. He did feel it should be a requirement that others cite the data set if they were to use it in their research.

Organization and description of data (incl. metadata)

Overview of data organization and description (metadata)

The researcher keeps a data dictionary that provides a description of the data. Typically the rows are people and the columns are variables.

Formal standards used

No formal standards have been used in the organization and description of this data. Despite this, the researcher felt the amount of organization and description is sufficient for another person with similar expertise to understand and properly use the data.

Locally developed standards

The data dictionary provides the primary means of description.

Crosswalks

Not discussed.

Documentation of data organization/description

The researcher indicated it was a medium priority for him to make the data accessible in multiple formats and a low priority to apply standardized metadata from his field or discipline to the dataset.

Ingest / Transfer

Before the data could be ingested into a repository it would need to be de-identified for participants. It would also need to be cleaned, the data dictionary included and a project description written. The researcher would also want to include information regarding the methods for obtaining the data.

He indicated it he would want to include a statement regarding permissions and acknowledgments regarding data reuse.

The researcher indicated it would be a medium priority for him to submit the data to a repository himself and to have the ability to easily transfer the data to a permanent data archive.

It was not a priority for him to make the process of submitting data be automated and he didn't know the priority for batch uploading the data.

Sharing & Access

Willingness / Motivations to share

The scientist was only willing to share data in the interim complete stage and the final complete stage.

He was willing to share the interim complete data only with immediate collaborators. He was willing to share in the final stage, incrementally to others after completion of all analyses by project stakeholders.

At this time he is unsure if he would be willing to submit his data to a data repository. He was interested in learning about the goals and purpose of the repository.

Embargo

No embargo period was discussed.

Access control

The scientist listed the ability to restrict access to the data set to authorized individuals as a medium priority. He was interested in somehow restricting access to WUSTL only students and researchers.

Secondary (Mirror) site

The researchers found the ability to access the data set at a secondary site as a low priority.

Discovery

The scientist found it to be a medium priority for researchers within and outside his discipline to find his data set. It was a low priority for the general public to find his data set and a medium priority for the data set to be discoverable through Internet search engines.

Ideally, he would like to partner with capable people who are interested in the data and work with them to find the information or collaborate.

Tools

Most of the data was collected via phone interviews and was self-reported. There is some data that is automatically generated by the 211 phone system, including average call duration and time in queue.

To utilize the data a number of software programs are used, including SPSS, SAS, Excel, and Filemaker software. They've also developed their own program that pulls down data from the 211 telephone system.

The scientist felt it was a low priority to connect the data set with visualization or analytical tools and a low priority for others to have the ability to comment on or annotate the data set.

Linking / Interoperability

It was a high priority for him to have the ability to connect or merge his data sets with other data sets.

It was a medium priority to support the use of web services/API and a low priority to connect his data with other publications or other outputs.

Measuring Impact

Usage statistics & other identified metrics

It was a medium priority for the faculty member to be able to see usage statistics on how many people have accessed the data, gather information about the people who have accessed the data, and have the ability to track data citations.

It was a low priority to track and show user comments.

Gathering information about users

He would be interested in a metric that would show conversion to scholarship. For example, how many uses of his data resulted in publications or other tangible scholarly products?

Data Management

Security / Back-ups

The scientist's data is currently stored on a local server and backed up daily. He has a firewall in place to protect his data.

Secondary storage sites

It was of medium priority to the scientist to have a secondary storage site and a secondary site at a different geographic location.

Version control

The researcher indicated it was a high priority to him to enable version control for this data set.

Preservation

Duration of preservation

The scientist believes the data will have value 3 years or more but less than 5 years from the time of completion.

Data provenance

It would be important to preserve the final core data set, the data dictionary and the boilerplate descriptions of the project.

Data audits

It was a medium priority for the faculty member to have the ability to audit his data set over time.

Format migration

The ability to migrate the data set to new formats over time is a medium priority for the scientist.