Data Curation Profile – Cornell University, Demographics

Profile Author	Keith Jenkins
Author's Institution	Cornell University
Contact	Keith Jenkins, kgj2@cornell.edu
Researcher(s) Interviewed	Withheld.
Researcher's Institution	Cornell University
Date of Creation	2012-03-22
Date of Last Update	2012-03-22
Version of the Tool	1.0 / modified
Version of the Content	1.0
Discipline / Sub-Discipline	Sociology / Demographics
Sources of Information	• An initial interview conducted on March 9, 2012. • A second interview conducted on March 13, 2012. • A worksheet completed by the scientist as a part of the interviews. • Project website, and sample meta-tables from internal database
Notes	
URL	http://www.datacurationprofiles.org http://hdl.handle.net/1813/29064
Licensing	

Section 1 - Brief summary of data curation needs

The scientists currently acquire demographic data from a variety of sources, then process, analyze, and aggregate the data, storing the results in an MS SQL database. A public website, developed by the scientists, provides open access to this final data in the form of charts, tables, maps, and downloadable Excel spreadsheets.

The current project website is quite successful, but it has required a lot of local development effort. The scientists seem quite interested in the idea of utilizing an external data repository, but would want to ensure that it would be able to offer web APIs similar to what they currently have in place, so that their public website would be able to draw data directly from the repository, for example.

Section 2 - Overview of the research

2.1 - Research area focus

The scientists take data from the US Census and other federal and state government sources, and add value to it by processing, analyzing, and distributing this data on their project website in a way that makes the data more accessible and easier to use. They also produce analytical

reports that are also available on the project website. The scope of this data curation profile is limited to the public data that is delivered via the project website.

2.2 - Intended audiences

The primary audience is the New York State Data Center and its affiliates. The data and reportrs are also used by other researchers, educators, and the general public.

2.3 - Funding sources

The project is funded by New York State's "Empire State Development" (ESD) -- formerly known as the Department of Economic Development. As recipients of this funding, the scientists produce and distribute data to the public, respond to questions from researchers and journalists, and produce estimates and projections as active members of the Federal-State Cooperative for Population Estimates (FSCPE), and the Federal-State Cooperative for Population Projections (FSCPP).

Section 3 - Data kinds and stages

3.1 - Data narrative

The first stage is the Acquisition of the source data, which consists primarily of public data from the US Census Bureau, but also some preliminary (non-public) Census data, and also some data from other sources such as the US Bureau of Labor Statistics (BLS), and New York State Dept. of Health.

The second stage is Appending new data to existing time series in the MS SQL database.

The third stage is the Aggregation of the data to state economic regions, to support the production of reports for each region.

The fourth stage is Estimates/Projections, which are calculated and reported back to the Census.

The fifth stage is the Website, where users can view charts, maps, and tables that are dynamically created via an automated process that pulls data directly from the MS SQL database.

3.2 – The data table

Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
Primary Data				
Acquisition	Original data from government sources	25 annual files, 15 decennial files < 500 KB, except for shapefiles (which may be ~100 MB)	.xls, .csv, .html, .shp, fixed-width text files	Public data and restricted data from US Census, BLS, NY Dept. of Health, etc.
Append	Annual data appended to existing time series		MS SQL	
Aggregation	Data summary by region			Reported to ESD
Mapping	Static maps	~150 static maps ~100 KB each	.jpg	Created using ArcGIS
Estimates/ Projections	Calculated values			Reported to ESD and Census (FSCPE, FSCPP)

Website	Data for public access		.html (with embedded tables, charts, trendlines, maps), .xls	Tools: ColdFusion, Google Chart API, Google Map API

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

3.3. - Target data for sharing

Most of the data can be shared with the public. The exceptions are preliminary, embargoed data that the Census provides for review by the scientists before being publically released by the Census. Some of this restricted data cannot even be sharing among other members of the project.

3.4 - Value of the data

The data is of value to a wide range of researchers in many fields, but especially policymakers, planners, school administrators, and news media who are interested in using demographic data and examining trends over time. The scientists package the data in a way that is particularly useful for those groups.

3.5 - Contextual narrative

Most of the data begins and ends as public data. Once the data has been processed and loaded into their MS SQL Server, it can then be queried and visualized on the project website.

Section 4 - Intellectual property context and information

4.1 - Data owner(s)

Most of the data originates from the US federal government, and is in the public domain. Although the scientists develop estimates and projections at the request of the state, these estimates and projections are considered to be owned by the scientists' institution, and are also made freely available to the public.

4.2 - Stakeholders

The data stakeholders include the funding source, Empire State Development (ESD) as well as the US Census, to whom the scientists report some of their aggregated, estimated, and projected data. But beyond some basic commitments to these stakeholders, the scientists are free to pursue their own research and further development of the data.

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

For the public data (the majority), no specific terms of use were discussed. The few restricted data files would not be published or redistributed at all.

4.4 - Attribution

The scientists do not require attribution for the use of their data, but would appreciate any acknowledgement.

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

5.1 - Overview of data organization and description (metadata)

The data is stored in several tables in an MS SQL database, which also includes some "metatables" that describe the original source of various tables and variables. These metatables also include configuration information for the public website, such as short and long names for variables, numeric format, colors for mapping, etc.

5.2 - Formal standards used

The data uses several standard Census variables, and, as is typically the case with Census data, one would need to consult external Census documentation for complete variable definitions.

5.3 - Locally developed standards

None

5.4 - Crosswalks

None

5.5 - Documentation of data organization/description

In addition to the "metatables" mentioned above (5.1), the project website also contains some documentation (such as "Methodology description") in PDF format.

Section 6 - Ingest / Transfer

The scientists would be willing to transfer their data to an external repository, after initial processing and documentation of the data. They would like to be able to initiate the data submission themselves. Batch processing was not seen to be a priority, nor was the ability to transfer to a permanent data archive, since most of the source data is already publically available. (However, see section 13 regarding the need to preserve the original source data.)

Section 7 – Sharing & Access

7.1 - Willingness / Motivations to share

The scientists are willing to submit data to an open access data repository, once the initial data has been processed. Enhancing public access to this data is one of the main goals of the scientists.

7.2 - Embargo

The scientists did not see any need to embargo their data.

7.3 - Access control

The scientists were interested in the ability of a repository to restrict access to specific individuals. Although most of their data is public, such an ability could conceivably be useful for handling preliminary estimates, or other restricted data.

7.4 Secondary (Mirror) site

Mirroring was considered a low priority.

Section 8 - Discovery

The scientists place a high priority on the discoverability of their data. Due to the nature of the data, this applies equally to researchers both within and outside the discipline of demographics, as well as the general public.

Users currently find the project website via Google and other search engines, and website log analysis shows that some users are searching for the project name, while others are searching for terms like "county population" and "projections". Other users come to the project website by following links from referring websites, including state and county government websites.

Section 9 - Tools

The scientists use MS SQL, SAS, and Excel to process the data. The project website uses ColdFusion to query and retrieve and display data from the MS SQL database. The website also uses the Google Maps API to display dynamic maps of the data, and the Google Charts API to visualize data in various forms, such as population pyramids and trendlines. Data can also be downloaded from the website as Excel spreadsheets.

If this data were hosted in an external data repository, it would be a high priority to be able to continue to use visualization tools such as Google Maps and Google Charts. The ability to annotate or comment on the data set was a lesser (medium) priority; possible uses would include knowing when a user reports that more recent data is available elsewhere.

Section 10 – Linking / Interoperability

Most of the scientists' reports are published on their project website. The ability to connect the data directly to the publications was viewed as a low priority. However, the availability of web APIs was a high priority – they have already created their own internal APIs to support their website.

Merging their data with other state or county-level data was also considered a high priority.

Section 11 - Measuring Impact

11.1 - Usage statistics & other identified metrics

Tracking usage statistics is a medium priority. Google Analytics is currently used to track pageviews on the project website. Tracking data citations was also a medium priority, but the ability to track user comments was marked as a high priority.

11.2 - Gathering information about users

The scientists are most interested in knowing the most frequent Internet domains of their users, and the referring keywords that were used to find the website.

Section 12 – Data Management

12.1 - Security / Back-ups

Backups of the data (on staff computers and the web server) are managed according to the standard practices of the college's IT department. Any restricted, non-public data is stored on CRADC (Cornell Restricted Access Data Center).

12.2 - Secondary storage sites

Secondary storage sites are not a priority for this data.

12.3 - Version control

The scientists ranked versioning as a low priority.

Section 13 - Preservation

The scientists stated that it would be most important to preserve the data as acquired from the original sources (as opposed to their end product), since there are no clear assurances that those original sources would still have the same exact data snapshots available in the future. (Some sources may only provide current data, but not data series years into the past.)

13.1 - Duration of preservation

The scientists felt that this data would be useful for between 20 to 50 years. One of the main objectives of their project is to present the data as a time series, so it is important to preserve a few decades of data, but after 50 years the data is “not that important anymore”.

13.2 - Data provenance

The scientists placed a high priority on documenting any changes made to the dataset over time – not necessarily full-fledged version control, but more in the sense of knowing whether any changes had been made, and by whom.

13.3 - Data audits

The scientists stated that they did not know how to rank the ability to audit the data over time.

13.4 - Format migration

The ability to migrate the data to new formats over time was ranked as a low priority.

Section 14 – Personnel

This section is to be used to document roles and responsibilities of the people involved in the stewardship of this data. For this particular profile, information was gathered as a part of a study directed by human subject guidelines and therefore we are not able to populate the fields in this section.