Data Curation Profile – Traffic Flow

Profile Author	J. Carlson
Institution Name	Purdue University
Contact	J. Carlson, jrcarlso@purdue.edu
Date of Creation	October 27, 2009
Date of Last Update	
Version	1.0
Discipline / Sub-Discipline	
Purpose	Data Curation Profiles are designed to capture requirements for specific data generated by a single scientist or scholar as articulated by the scientist him or herself. They are also intended to enable librarians and others to make informed decisions in working with data of this form, from this research area or sub-discipline. Data Curation Profiles employ a standardized set of fields to enable comparison; however, they are designed to be flexible enough for use in any domain or discipline.
Context	A profile is based on the reported needs and preferences for these data. They are derived from several kinds of information, including interview and document data, disciplinary materials, and standards documentation.
Sources of Information	• An initial interview with the scientist conducted in May 2008. • A second interview with the scientist conducted in December 2008. • A questionnaire completed by the scientist as a part of the second interview. • A White Paper/Report written by the scientist on the database design for this project
Scope Note	The scope of individual profiles will vary, based on the author's and participating researcher's background, experiences, and knowledge, as well as the materials available for analysis.
Editorial Note	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.
Author's Note	This Traffic Flow data curation profile is based on analysis of interview and document data, collected from a researcher working in this research area or sub-discipline. Some sub-sections of the profile were left blank; this occurs when there was no relevant data in the interview or available documents used to construct this profile.
URL	http://www.datacurationprofiles.org

http://dx.doi.org/10.5703/1288284315016

Brief summary of data curation needs

The scientist is seeking what he referred to as "Database Management Services". In addition to IT support for managing, hosting, and staff training on databases, the scientist would like assistance in developing sound data management and documentation practices. The scientist repeatedly mentioned the need for developing and implementing a standardized workflow for his data so that it is structured and described in a manner that enables it to be easily discovered, accessed and used by him or re-used by others, as well as ingested and archived with a data repository.

The scientist needs to be able to correlate the different data kinds together: sensor data, images, and possibly 3rd party data (weather and road conditions). Some of his data lives in an SQL database and some of his data lives in Excel spreadsheets.

The scientist needs to be able to share his data with his collaborators at other institutions electronically before making this data publicly available.

Use of the data would require that attribution be given to the scientist.

The data would benefit by visualization tools to enable/enhance its use. Ideally, the user of the data would be able to generate his/her own charts, graphs, maps and/or other visualizations.

Overview of the research

Research area focus

The scientist studies real-time traffic signal performance measures in which he measures the movement of traffic, specifically the number of vehicles passing through an intersection and the amount of time they spend at an intersection on a movement-by-movement basis over a 24 hour period. The result is a profile of traffic movement for an intersection. The scientist described the project's aim as enabling INDOT to make data-driven decisions to prioritize which traffic signals they should work to improve. This research and the resulting algorithms are also of interest to vendors of sensor equipment and could be used to improve their performance.

Intended audiences

- Civil Engineers
- Other researchers studying transportation and traffic flow issues.
- State Government – esp. Departments of Transportation
- Industry – sensor manufacturers
- General Public

Funding sources

The Indiana Department of Transportation (INDOT) and the National Cooperative Highway Research Program (NCHRP) are the primary funders. The National Academy of Science was also mentioned as a funding agency for this project. Other secondary funders include private sector companies, the United States Department of Transportation (USDOT), and a local technical assistance program.

The scientist has not been mandated by his sources of funding to generate a data management plan or share his data with others outside of his lab. He reports that none of the funding sources have had an interest in the data itself from this particular project thus far. Their interest is primarily in the resulting algorithms and models for determining and measuring traffic flow.

The scientist did mention that the potential value of the data to INDOT and perhaps some of the other funding agencies would grow if this research were to be increased in scale and gathered on

a systematic basis. The data could then be used as a basis to support "integrated corridor management."

Data kinds and stages

Data narrative

The scientist and his research collaborators have placed sensors and video cameras to monitor traffic flow at several intersections around Indiana. The road sensors are placed in each lane of traffic. The frequency of the sensor changes when a car or other vehicle passes over it and then reverts when the vehicle rolls off of it. Sensors in the traffic light record the status of the intersection (that the light is red, yellow, or green). Data from the sensors are FTP-ed out on an hourly basis as compressed files. Data are then processed, normalized and reformatted from the vendor's proprietary format into .csv and then into Microsoft Excel.

Video of the sites are taken for data verification purposes. The video gathered is parsed out into .gif or .jpg images at the rate of 20 frames per second.

The data and corresponding images are then imported into a mySQL database approximately every 2-3 weeks. Data are extracted from the mySQL database for analysis purposes back into Microsoft Excel. Excel is used to generate pivot tables, charts and graphs. Selected pivot tables are placed into power point slides for the purposes of presenting the data and research findings to others.

The scientist gathers data on weather conditions from the "Weather Underground" website for explanatory purposes as weather conditions may affect traffic flows. The scientist also occasionally gathers data from INDOT databases regarding road conditions for explanatory purposes. It is not clear precisely when or how this data are correlated to the traffic flow data described above.

The categories in the "data stage" column listed in the table below were developed by the authors of this data curation profile. The data specifically designated by the scientist to make publicly available are indicated by the rows shaded in gray.

Data Stage	Output	Typical File Size	Format	Other / Notes
Primary Data				
Raw	Sensor data	100k in 1 file per day	The format is proprietary to the sensor	FTP downloads are mostly automated.
Processing Stage 1	Sensor data – normalized, screened for outliers/errors, and moved to an open/accessible format	Roughly 6kb	.csv / .xls	Data are formatted into .csv before bring reformatted into a mySQL database.
Processed	Data vectors	800 records per intersection per day. Each record has about 38 fields (floating point)	SQL / .xls	Data are extracted from the mySQL database for analysis purposes. The database typically holds 3-4 months worth of vehicle signature data, traffic signal data and the corresponding images.

Analyzed	Pivot charts/graphs		.xls / .emf	Data are placed into charts and graphs for interpretation. Visualization is needed to give data meaning and for presentation.
Published	Pivot charts/graphs		.ppt	Data are presented to others (incl. funders) via power point.
Augmentative Data				
Video			Several formats – primarily "Real Video" but .wmv, .mpeg as well	Video taken are correlated with the data for verification purposes.
Image	Stills taken from the video		.gif /.jpg / .ppt	Images are generated as still shots from the video.
3rd Party Data - Weather information from "Weather Underground" website			.csv files	Collected via screen scrape. Correlated with collected data for explanatory/ descriptive purposes.
3rd party data – road conditions from INDOT's databases			unknown	Collected on an ad hoc basis as needed for explanatory/descriptive purposes.

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.

Target data for sharing

The ingest package would consist of the spreadsheets containing the processed vector data from the sensor and the corresponding images of vehicles that pass through the site.

In addition, the weather data gathered from "weather underground," a 3rd party source, may need to be provided along side of the data as well. If this data were not ingested directly, there may need to be a link between the scientist's data and weather data. It is unclear if substituting weather data from a more authoritative source would be acceptable.

Data from INDOT databases may also need to be linked in somehow.

Use/re-use value of the data

This data will likely be valuable to different audiences in different ways:

- This data will help other engineers evaluate alternative loop-based vehicle re-identification schemes.
- The data could be used for longitudinal studies on traffic flow and congestion issues.
- The general public may be interested in the performance measures of traffic signals being generated through this data. The charts and graphs generated from this data demonstrate the traffic flow at an intersection at different times of the day.
- This type of data may be of interest to state agencies to better manage traffic flow on state highways. However, the data have to be gathered systematically from additional sites and

scaled up in order for it to be used in this manner. The tools needed for state to use this data effectively do not yet exist, though the USDOT is working on them.

Contextual narrative

The data are considered dynamic as they are still being generated and the data set continues to grow. The scientist was unable to provide an estimation of the eventual size of the data sets.

The data are used to test algorithms developed by the scientist and his collaborators.

Data are imported from the sensor in a proprietary format, cleaned up and normalized and then exported as a .csv file. It is then converted into an Excel spreadsheet. Some data are then imported into a mySQL database, primarily for storage. Much, if not all, of the data analysis is done through Excel. Ideally, for the purposes of the scientist, all of the data would be kept in a database not only for storage but also for analysis. However this transition has not yet occurred.

The scientist and colleagues have written up their database design and workflows in a white paper prepared for the NCHRP.

Intellectual property context and information

Data owner(s)

The scientist and his colleagues in this project see themselves as owning the data.

Stakeholders

The scientist sees funding agencies as having a stake in the data. The scientist stated that the funding agencies (INDOT, NCHRP and the vendors they are working with) want the intellectual property rights to the systems that result from the research; they do not want the data itself (at least at this point in time). Their interest is with the algorithm and/or the systems being produced through the research. The scientist did mention that he seeks out permission from the sponsor before making any data available before publication.

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

Attribution for use of the data must be given to the scientist.

Attribution

Receiving attribution from others who use his data in a manner that would count towards promotion and tenure considerations (being able to cite the data at minimum) is a high priority for the scientist.

Additional notes on intellectual property

There may be some privacy concerns with making the images used to verify the data from the sensors publicly available. The images identify the make and color of vehicles. However, the scientist believes that people or characteristics that would allow someone to identify people in the vehicle, such a license plate, are not discernable beyond a reasonable doubt. The risk is seen as negligible by the scientist, but further investigation on this issue would be required before ingesting the images into the repository.

Organization and description of data for ingest (incl. metadata)

Overview of data organization and description (metadata)

The data files are not as organized as the scientist would like them to be. However, he also indicated that the data are described well enough to be understood by others in his field. Currently, the data files are primarily organized by date.

The temporal and spatial relationships between the data are important; however it is unclear if the data are currently geo-coded or not.

A problem the scientist has encountered is the lack of standardized methods and procedures for organizing and managing his data. In previous conversations with the Libraries the scientist has expressed interest in persistence and naming authority.

Formal standards used

None. The scientist is not aware of any formal standards for traffic data, but assigned a high priority to adopting a formal metadata standard if/when one is developed.

Locally developed standards

The metadata are stored alongside of the data in the mySQL database (see "documentation of data organization/description" below). Standardized codes developed in-house to describe the status of the traffic signal.

Crosswalks

Not discussed.

Documentation of data organization/description

The scientist and his collaborators are considering developing a relational database to house and disseminate traffic flow data. A white paper describing the design of the relational database includes metadata tables. These tables are listed as follows:

- Sensors – sensor id, device id, lane id, ch, label, lat, lon
- Sets – set id, sensor id, start time, stop time, sensitivity, oversampling
- Devices – device id, type, vendor, model, serial
- Lanes – lane id, asset id, direction, label, phase
- Assets – asset id, lat, lon, description, is intersection
- State Codes – state code id, description

Definitions of each element listed are included in the white paper, although these definitions are very brief and more explanation may be required for curation purposes.

Ingest

The scientist has stated that not all of his data need to be ingested into a repository and archived. A selection and appraisal process and criteria will need to be developed.

The scientist's responses to the questions on submitting data to a repository (by one's self or through automated means) indicate that determining the means of ingest is a low priority for him.

Access

Willingness / Motivations to share

The scientist is willing (even eager) to make his data publicly available for a variety of potential uses on the condition that he receives credit, through a citation, when the data are used. He is very interested in employing DOIs for his data to enable their persistence so that the data may be cited.

Currently the majority of the scientist's data sharing activity is with INDOT and his collaborators. The scientist also shares the charts and graphs made from the data at professional conferences.

The scientist has found it very difficult to share data with collaborators outside of the university. A primary sticking point in sharing data with collaborators has been getting through university firewalls and getting people credentials to get through these firewalls. He is seeking ways to make data sharing with collaborators at other institutions easier.

Embargo

An embargo is needed only to give the scientist and his colleagues a chance to clean up the data and make it suitable for availability outside of the lab group. He estimates that this process will ideally take only 1-3 months, but may take as long as 1 year.

Access control

The ability to restrict access to the data to authorized individuals while the data is being "cleaned up" for public access is a medium priority for the scientist. Once the data has been cleaned the data may be released for public access.

Secondary (Mirror) site

The ability to access the dataset at a secondary site if the repository is "off-line" is a low priority for the scientist.

Discovery

The ability for researchers both within and outside of his discipline to find this data easily is a high priority for the scientist. The scientist also expressed a strong desire to make this data accessible to the general public through libraries or other means in a meaningful format (i.e. through Google maps or some other visualization interface).

Useful search functionalities within a repository mentioned by the scientist was the ability to find data by the date on which it was gathered and the ability to retrieve relevant weather data in conjunction with the traffic flow data. There may be a need to be able to search for the traffic flow data according to weather condition (e.g., query all data gathered on days with more than two inches of snow fell) though this was not mentioned explicitly by the scientist.

Tools

MS Excel (or csv reader) and image viewers would be needed to use this data.

Visualizations of the spatial and temporal elements of the data are an important component of its value and a high priority to the scientist. The users of the data should be able to generate graphs or charts from the temporal or spatial aspects of the data. Ideally, users of the data would be able to plug it into mapping applications or software. The data may need to be geo-coded in order to take advantage of such tools.

Interoperability

The scientist's data are composed of multiple components: sensor data (numeric) and images. The numeric data must be connected with the images that correspond to it according to their time stamps.

The scientist also makes use of 3rd party data are gathered from the "weather underground" website in order to explain certain differences seen in traffic flow on certain days. If this 3rd party data are not deposited into the repository alongside of the data generated by the scientist, there may need to be a link the repository to the "weather underground" site or another site providing

this type of information in a manner that would enable the same kind of functionality. The scientist has also used data from INDOT to verify work zone / construction incidents and so some kind of linkage between the repository data and INDOT's database may be desirable as well.

Developing connections between the data and any publications that have resulted from the data is a high priority for the scientist.

The ability for the repository to support the use of web services APIs is a medium priority for the scientist.

Measuring Impact

The repository should provide ways and means to measure impact that fit in with any and all evolving promotion and tenure considerations. This includes but is not limited to being able to cite the data. Having a Science Citation Index like functionality to determine its "impact factor" was also brought up by the scientist as a useful function.

Usage statistics

The ability to see usage statistics on how many people have accessed his data are a high priority for the scientist.

Gathering information about users

Gathering information about the users of his data was not discussed by the scientist.

Data Management

Security / Back-ups

Currently the scientist keeps his critical data on servers provided by his college. His understanding is that his data are kept in perpetuity on these servers. Data that is "less critical" is kept on two 4TB Buffalo servers, which mirror each other and are updated once a week.

Secondary storage sites

Secondary storage sites are not a priority for this data according to the scientist.

Preservation

Duration of preservation

The scientist believes that the data should be preserved for 3-5 years. The data may not be reliable beyond this time frame as road construction and other events may impact traffic flows.

Documentation of any changes that were (or would be) made to the data over time is a high priority for the scientist.

Data provenance

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

Data audits

The ability to audit the data to ensure its structural integrity is a high priority for the scientist.

Version control

Currently version control is not applicable to this dataset.

Format migration

Given the short term duration of preservation actions needed for this data, format migration is a low priority for the scientist.

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.

Primary data contact (data author or designate)

Data steward (ex. library / archive personnel)

Campus IT contact

Other contacts

Notes on Personnel

The scientist works with several graduate students on this project that do a considerable amount of work in gathering, processing and managing the data. They may need to be involved in the development of curation processes and workflows as well.