Data Curation Profile – Astronomy / Galactic Structure

Profile Author	Kathryn M. Dunn
Author's Institution	Rensselaer Polytechnic Institute
Contact	dunnk2@rpi.edu
Researcher(s) Interviewed	[Name withheld], Professor, Department of Physics, Applied Physics, and Astronomy
Researcher's Institution	Rensselaer Polytechnic Institute
Date of Creation	2012-09-21
Date of Last Update	2013-03-14
Version of the Tool	1.0
Version of the Content	1.5
Discipline / Sub-Discipline	Astronomy / Galactic Structure
Sources of Information	• An initial interview conducted on May 29, 2012. • A second interview conducted on May 30, 2012. • A worksheet completed by the researcher and interviewer as a part of the interviews. • A published article based on the data described in the data curation profile. • The researcher's webpage • Follow-up emails from the researcher, received January 18 and February 28, 2013.
Notes	Three additional local questions were added to the interview. See Section 15 – Local Questions.
URL	http://dx.doi.org/10.5703/1288284315058
Licensing	Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License. cc BY NC SA

Section 1 - Brief summary of data curation needs

The researcher participates in large-scale astronomical survey projects, which are collaborative efforts of many institutions and individuals. Management of the raw and processed/calibrated data is handled by those projects. The researcher would like to have a better way to share the data behind the figures in their published papers with other researchers. The effort involved in retrieving and documenting the data so it's usable by others is a barrier to sharing the data. It would be easier to share data from publications if it could be deposited to a centralized repository (institutional or elsewhere), where it could be maintained past the length of a grant-funded project. It would be very important for the data to be linked from the article in ADS (Astrophysics Data System - http://adswww.harvard.edu/), so that other researchers would know it was available.

Section 2 - Overview of the research

2.1 - Research area focus

The researcher's current research focuses on the formation and structure of the Milky Way, including the distribution of dark matter in the galaxy. The researcher uses different types of stars to trace the structure and evolution of the Galactic halo and disks.

The specific research project detailed in this data curation profile was completed in 2002, but the researcher said that it is still representative of how they conduct research and work with data today.

Researcher: "This is one of the first projects that was done with the Sloan Digital Sky Survey data. We tried for some time to fit a smooth density model to the stars in the spheroid of the galaxy and didn't have much luck, depending on which slice of data we used. And finally when we plotted up where the densities of stars were, we noticed that there were areas where there was a higher density of stars just in certain parts of the sky, and the reason that we were unable to fit the models is that the models didn't have this strong overdensity of stars in one place. This was the first time that stellar substructure in the Galactic, the stellar halo was discovered through density. Before this, people had thought that you'd see substructure where you'd have a whole bunch of stars that came into the galaxy in one lump, and they would all be moving together, but no one realized that they would stand out in density [instead of just velocity]."

Researcher: "This was kind of the discovery that the spheroid of the Milky Way was lumpy in density, in a significant way."

Researcher: "At the time that the data was taken, the telescope couldn't even move, it was just fixed on the crater, so it's on the celestial equator. It's a strip of sky that is 2 and a half degrees wide in declination and […] you have two sections of more than a hundred degrees in length, one on the north Galactic cap and one on the south Galactic cap. And the Sloan Digital Sky Survey data, there's images, but the survey itself processes those images to derive the parameters for each object in the image, and so we were working not from the image data, but from the derived parameters from the images. And so for each object, we get the magnitude, or the brightness of the object, in each of five different filters, u, g, r, i, and z, and we can tell which of the stars is a point source, like a star or a quasar, and which ones are galaxies. For this study, we selected out just the ones that were point sources and we also, for the most of the study, selected out just the ones in a narrow g- r color range so we could pick out a particular type of star."

2.2 - Intended audiences

Other researchers in the same field.

2.3 - Funding sources

The specific research project described by this project was funded by NSF. At this point, NSF did not require a data management plan, but it does now (see Section 15 – Local Questions for the researcher's perspective on DMPs).

Sloan Digital Sky Survey (SDSS), the project that is the source of the data, is funded by the Sloan Foundation, NASA, the NSF, the Department of Energy, and each of the member institutions of SDSS.

Section 3 - Data kinds and stages

3.1 - Data narrative

The initial data stage (1) is observations collected from the SDSS telescope. In the next data stages (2a, 2b, 3a, 3b), the observations are then analyzed and calibrated through a set of complex data pipelines at SDSS. These initial data stages are managed and maintained by SDSS. The science use database (3b) is made available in a series of public data releases (though this project used pre-released data). In the next stage, the researcher uses SQL queries (4a) to query the science use database (3b). This results in a large ASCII table (4b) which describes the subset of objects (~4.3 million stars) and parameters which are relevant to the researcher's research questions. In the next data stage, this large ASCII table is subsetted by a series of scripts (5a) to investigate specific research questions. This results in directories of subsets of the large ASCII table (5b). The researcher noted that it's more important to maintain the scripts that subset the data than the subsetted data itself, which can always be regenerated. Additional scripts (6a) are used to generate the figures and tables for publication, which are considered the final data stage (6b). A few of the published figures are edited manually for publication (to add axes, etc.), but most of the figures are generated in their entirety by scripts.

3.2 – The data table

Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
Primary Data				
1. Telescope observations	Original observation images from telescope, written on tape	Many hundreds of files. Size from 32 MB down to next to nothing	FITS (Flexible Image Transport System) images and a few ASCII files	Maintained by SDSS
2a. Data analysis pipelines	Pipelines used to convert telescope observations (1) to operational database (2b)		TCL, C	Maintained by SDSS
2b. Operational database	Objects and their parameters after image data has gone through Sloan data pipelines, postage stamp images		FITS binary tables (in a Sloan-specific FITS format) and some ASCII	Maintained by SDSS
3a. Calibration pipelines	Pipelines used to convert operational database (2b) to science use database (3b)		TCL, C	Maintained by SDSS
3b. Science use database (Analyzed and calibrated data)	Calibrated and annotated data that has been optimized for rapid access by a larger community	Hundreds Some are 8-16 MB, catalogs are 800 kb	Catalog Archive Server Jobs System (CasJobs), can be queried online using SQL	Maintained by SDSS, made available in public data releases (though this project predates the public releases). Continually reprocessed/recalibrated and rereleased.
4a. SQL queries	Queries used to retrieve data from SDSS science use database (3b) to create "Big ASCII table" (4b)	Less than 5 queries	SQL	Researcher writes these queries in order to retrieve just the objects and parameters that are useful for their research question.

4b. “Big ASCII table” - Subset of objects and parameters	Big ASCII table with subset of objects and parameters that are useful to researcher	One large table with one line for each of ~4.3 million stars. Size on order of GBs.	ASCII	Generated by researcher from Sloan data. Result of SQL queries (4a) on Sloan science use database (3b).
5a. TCL/awk scripts to subset the table	Scripts to subset big ASCII table, enrich with additional information and generate figures	At least 1 script per figure in paper.	TCL and AWK scripts	More important to preserve the scripts (5a) than the data they generate (5b) – data takes up a lot of space and can always be regenerated.
5b.Directory of subsets of big ASCII table		More than 1000 files. Size varies, 1000’s of MB – 2 million F turnoff stars	ASCII files	More important to preserve the scripts (5a) than the data they generate (5b) – data takes up a lot of space and can always be regenerated.
6a.TCL scripts that generate published figures and LaTeX tables		At least 1 script per figure in paper	TCL scripts	
6b. PostScript figures and LaTeX tables for publication	Published figures as appearing in the journal	26 figures Not sure about size.	Encapsulated PostScript, LaTeX	Figures are created from TCL scripts (6a), sometimes with some hand edits.

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 researcher did not provide a response.

3.3. - Target data for sharing

In principle, the researcher would be willing to share anything after publication, as long as the project collaboration allows it, but sometimes it's too much work to get the documentation together, or it takes a long time to get around to it.

At the time this paper was published, none of the data could be shared because of the rules of the SDSS collaboration, but versions of stage 3b have since been released and are shared publically by SDSS.

The data and scripts used to generate the final published figures (5b and 6a) would be most valuable to other researchers.

3.4 - Value of the data

Many people already use the stage 3b data, but that is already managed and shared by SDSS. The data and scripts from stages 5b and 6a would be useful to researchers who would like to build on this specific published research by overlaying new data on the published images or creating a new inset. Currently, to accomplish this, researchers add data to the final PostScript images (stage 6b), but it's difficult to manipulate those. The researcher also noted that it would be easier to find a particular data point in someone else's figure while reading the article if the data were readily available – people typically estimate locations using a ruler on the printed journal article, or by reading the PostScript file.

3.5 - Contextual narrative

The researcher discussed the Astrophysical Journal's plans to support distribution of data behind published figures. The researcher has heard that Astrophysical Journal is interested in this, but is not sure what the status is.

Section 4 - Intellectual property context and information

4.1 - Data owner(s)

Stages 1-3:	The SDSS organization
Stages 4a-6a:	paper authors
Stage 6b:	journal

4.2 - Stakeholders

The researchers participating in the SDSS collaboration and the journal publisher.

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

In principle, the researcher would be willing to share anything after publication, as long as the project collaboration allows it. If someone were to reuse or modify a published figure, they would need to contact the researcher for permission and cite the source paper that contained the figure, in any subsequent works produced.

4.4 - Attribution

Typically in astronomy, when someone else re-uses or modifies figures, they will contact you for permission and cite the paper that they got the figure from in their paper. The researcher prefers that they cite the paper that the data came from, rather than the data itself.

Sometimes if one researcher works with another researcher's data, they will be co-authors on the paper.

Sloan itself also has a Scientific and Technical Publication Policy, which describes "the policies and guidelines governing the publication of scientific and technical results from the Sloan Digital Sky Survey": http://www.sdss.org/policies/pub_policy.html

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

5.1 - Overview of data organization and description (metadata)

Stages 1-3: Outside the scope of this Data Curation Profile. They are managed and documented by Sloan and hugely complicated. There is lots of documentation, but the data are always being updated and the documentation is never enough.

Stages 4a-6a: ASCII files with a README file in the directory that says what the column headings are. Data used to make the figures is documented by the scripts that make the figures. Existing documentation may or may not be sufficient for another researcher in this field to make use of it. But, even if other people can't run the scripts in their local environment, they can look at them to see what was done to make the figures.

Stage 6b (figures and tables): documented in journal article

5.2 - Formal standards used

Stage 1 is in FITS image format. Stage 2b and 3b are in a SDSS-specific FITS binary format. Field-specific standard definitions of coordinates, magnitudes, etc. are used.

5.3 - Locally developed standards

Did not discuss.

5.4 - Crosswalks

Did not discuss.

5.5 - Documentation of data organization/description

Stages 1-3 (for the various data releases – the data in this paper preceded data release 1) are documented through the SDSS website (http://www.sdss.org) and through various published papers on the collection and processing of the data (SDSS Technical Publication List - http://www2.astro.psu.edu/users/dps/sdsstechrefs.html).

Stages 4a-6a are documented by README files stored on the directory with the scripts and data.

The published paper itself documents the data in a broad sense, especially stage 6b (the figures and tables included in that paper).

Section 6 - Ingest / Transfer

The researcher noted that additional work would be required to document data stages 5b and 6a so that they would be more usable by others before depositing it in a repository.

Physical transfer of the data was not discussed.

Section 7 – Sharing & Access

7.1 - Willingness / Motivations to share

In principle, the researcher would be willing to share anything after publication, as long as the project collaboration allows it. The researcher says that they are willing to share their data in principle, but are concerned about the time and effort it would take to get the data into a state where it would be useful to other researchers. If time and effort were no object, the researcher would be willing to share any component of the data.

7.2 - Embargo

Embargo may be required, depending on the rules of the collaboration. (This data was not able to be shared at the time of publication – it preceded the SDSS public data releases.) The researcher would like to hold the data until publication of the associated article, but is fine with releasing all data at the time of publication. The researcher does not want someone competing with them while they are still developing their results. The journal may also want to embargo the data.

7.3 - Access control

Access control is not a priority for the researcher.

7.4 Secondary (Mirror) site

High priority for stages 1-3, but this is managed by Sloan. Low priority for later data stages (those related to this specific research project).

Section 8 - Discovery

It's very important for other researchers in the field to be able to discover the data, otherwise managing and sharing the data is not worth doing. It doesn't matter where the data are stored, as long as it's linked to the article in Astrophysics Data System (This is an existing functionality in ADS.) If it's linked there, people will find it.

It's very important for the general public to be able to access stage (3b), but Sloan already makes it available. It's not important for general public to be able to discover other stages. Discovery by researchers in other disciplines is a low priority.

Discovery through Google and other search engines is not a priority.

Section 9 - Tools

The original data was generated by a telescope and CCD cameras that write to tape. The stage 2b and 3b data was processed by data pipelines (2a, 3a) written in TCL and C. Stage 4b data was pulled from the SDSS database using SQL queries (4a). Stage 5b data was generated by AWK scripts and TCL (5a).

To use the data you need any package that reads FITS images or ASCII tables.

The ability to connect the data set to visualization or analytical tools is extremely important for data stage 3, but not important at all for data stages 4b and 5b, because those are already visualized in the paper figures (stage 6b).

The ability of others to comment on or annotate the data set is a low priority.

Section 10 – Linking / Interoperability

The ability to connect the data with publications or other outputs is a high priority.

If people reuse data or figures, they should cite the paper that contained the figures or is associated with the data they used.

The ability to support the use of web services and APIs, and the ability to connect or merge the data with other data sets, is a high priority for stage 3b, but not a priority for stages 4-6.

Section 11 - Measuring Impact

11.1 - Usage statistics & other identified metrics

The ability to see usage statistics on how many people have accessed this data is a medium priority. This would be useful in demonstrating impact when applying for funding in the future.

It would also be useful to know the citations of any publications based on the data.

11.2 - Gathering information about users

The ability to gather information about who has accessed or made use of this data is a low priority.

Section 12 – Data Management

12.1 - Security / Back-ups

Stages 1-3 are managed through Sloan. For stages 4-6, the researcher maintains servers set up with RAID arrays / disk striping. Currently the researcher cannot read the RAID arrays with the non-public files for this paper, because the computer does not boot. The researcher has not tried to recover the disk because there is no one available to do that at this time. Important scripts are backed up through a CVS repository that is backed up to Fermilab. It is more important to back up the scripts (4a, 5a, 6a) used to generate the data in stages 4b, 5b, and 6b than the data itself. The data takes up a lot of space (TB / hundred GB) and can be regenerated using the stage 3b data if something catastrophic happens.

12.2 - Secondary storage sites

Important scripts are backed up through a CVS repository that is backed up to Fermilab.

12.3 - Version control

Version control is very important for the stage 3b data, which is managed by Sloan and is continually being recalibrated/rereleased. It is not a priority for the other data stages.

Researcher: "Well, you certainly need to document changes that were made to the dataset. I don't know that I would be thinking of making any, though."

Important scripts are version-controlled through a CVS repository that is backed up to Fermilab.

Section 13 - Preservation

The most important things to preserve are the original data (stages 1-3) and the final figures (stage 6b). These are already preserved through Sloan and the publisher of the journal article. The next priority for preservation would be the data and scripts used to generate the published figures (stages 5b and 6a). The other intermediary data between the original data and the data for the figures is the lowest priority.

A secondary storage site is only important if it's at a different geographic location.

13.1 - Duration of preservation

Preservation of stage 5b data used to generate figures: "If people are still citing the paper, then they probably still want the data from the figures." People are still citing this paper 10 years later, and the citation rate is as high as ever. Estimated duration 10-20 years.

13.2 - Data provenance

Not discussed.

13.3 - Data audits

Low priority for stages 4-6, especially if it requires the researcher's attention. "I don't have time for that."

13.4 - Format migration

The ability to migrate datasets to new formats over time is a medium priority.

Section 14 – Personnel

The data described in this research (stages4-6) are not being actively curated.

14.1 - Primary data contact (data author or designate)

(Withheld)

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

N/A

14.3 - Campus IT contact

N/A

14.4 - Other contacts

N/A

Section 15 – Local questions

The following questions were asked outside of the context of the data set that was used to generate Sections 1-14 in this DCP.

15.1 - Experience fulfilling the NSF Data Management Plan requirement

The NSF began requiring that all grant proposals include a data management plan as of January 18, 2011. Could you tell me about your experience with fulfilling the Data Management Plan requirement for your NSF grant proposals?

The researcher is knowledgeable about data management – participated in planning for the SDSS project, and participated in NSF panels for big data management grants.

The proposals that the researcher has submitted since the DMP requirement took effect have not involved the management of primary data, just the secondary data used for that particular project. Thus far, preserving this data through the end of the project using RAID and CVS has been sufficient for NSF's needs. The researcher feels confident in writing a DMP for data that is not going to be preserved past the life of the project, but would need help if they had to write a DMP for a project that generated a lot of primary data that needed long-term preservation.

Since the NSF DMP requirement is so new, the first grants to be given under that requirement have not yet gone back for renewal, so it's too soon to tell how/whether the NSF will incorporate checking that the DMP has been fulfilled as a condition of renewal.

Based on town hall meetings the funding agencies (including NSF) have given at the American Astronomical Society Meeting, the researcher believes that funding agencies are taking long-term preservation of primary data very seriously. Astronomers still use photographic plates from a hundred years ago, but from the mid-1980s (when data began to be collected digitally), there is a gap where the data may or may not still be readable. Example: the researcher's thesis project data are in a big box on a bookshelf. "Even if you had an exabyte drive, they're probably not readable."

15.2 - Need for assistance with data management or data management plans

What kind of assistance would you find helpful in managing your data or creating data management plans for funding agencies?

It would be helpful to have a centrally managed and backed-up repository (institutional or elsewhere). Currently there is no way for the researcher to preserve data past the end of a project. It would also be helpful to have a repository available to manage data and scripts used to create figures from published articles.

Currently many researchers share the data management plans they've created with others.

Based on their participation on NSF panels, the researcher believes that it's important for our institution to be one of those that have the facilities to preserve data, as the funding models for long-term preservation are developing. Eventually institutions who have long-term data preservation capabilities may charge depositors from outside the institution for these services.

The researcher feels confident in writing a DMP for data that is not going to be preserved past the life of the project, but would need help if they had to write a DMP for a project that generated a lot of primary data that needed long-term preservation.

15.3 - Contributing data to a pilot repository

The Tetherless World Constellation has a pilot data repository for the management and dissemination of Rensselaer research data running at http://data.rpi.edu. Depositing your data in data.rpi.edu makes it easier for others to find, use, and cite your data, and can help you meet the requirements of funding agency data management plans. Do you have any data you would be interested in making available through data.rpi.edu?

The researcher had already contributed some data to the repository, but, is interested in possibly depositing figure data and scripts (stage 4 data) from published papers on an ongoing basis. The researcher originally thought that the work involved in documenting the data enough to make it usable to others would be a barrier to depositing it in a repository. As a result of discussions during the Data Curation Profile interviews, they now think that "preserving the figure information could be pretty useful, actually, and not all that much work." In the past when people have requested data, they have sometimes sent them scripts, and even if they're unable to actually run them (don't know TCL or can't run it) the scripts document what was done to create the figures. People can use them as pseudocode to write their own scripts.