Data Curation Profile - Astrophysics

Profile Author	Ardys Kozbial
Institution Name	University of California, San Diego
Contact	Ardys Kozbial, akozbial@ucsd.edu
Date of Creation	October 21,2010
Date of Last Update	
Version	1.0
Discipline / Sub-Discipline	Astrophysics
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 scientist/scholar's 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 August, 2010. • A second interview with the scientist conducted in September, 2010. • A worksheet completed by the scientist as a part of the interviews. • A published paper explaining the research and the methodology used to gather, process and analyze the data set in question.
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 Astrophysics data curation profile is based on analysis of interviews, a completed worksheet and information gathered from a publication, 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/10.5703/1288284314996

Section 1 - Brief summary of data curation needs

The data set contains large scale simulations of the universe focusing on galaxy clusters and properties. It includes 28 snapshots of the universe and data generated at each of four stages: Simulation, Data Reduction, Analysis and Publication.

The data are currently posted on the scientist's web site after the results of his work have been published.

Ideally, the scientist would like to make the data available from a central repository to his collaborator throughout the four stages (see Section 3 for details). After publication, data from all stages can be made available more generally.

If there are not sufficient controls in the central repository to limit access to collaborators, he is content to share data with collaborators another way during the first three stages.

Section 2 -- Overview of the research

2.1 - Research area focus

The scientist describes his work in cosmology as large scale simulations of the universe focusing on galaxy clusters and properties. The researchers take snapshots of the universe, freezing them in time, and derive images from them. Each snapshot results in a separate data set and each data set can be analyzed multiple times.

One of the papers that resulted from his work, and the end product of this data set can be found at http://arxiv.org/abs/0704.2607v3.

2.2 - Intended audiences

Other astrophysicists will be interested and able to interpret data at all stages.

Science educators may use data from the later stages (movies generated, visualizations) for outreach in classes to show cosmology – star formation, turbulence.

2.3 - Funding sources

National Science Foundation (NSF)
NASA

At the time the research was being done, neither funding source required data management plans or proof of data sharing as a condition of funding.

Section 3 - Data kinds and stages

3.1 - Data narrative

The first stage is Simulation. This is a data stage that generates a set of initial conditions, a model of the universe at an early age. Differential equations are applied and the universe is evolved. Its state is captured at different times. The process sounds straightforward, but it is a lot of work because it is at a large scale. The purpose of this stage is to try to define the state of the universe. A model of the universe is evolved so that there is something to go back to and measure.

The second stage is Data Reduction. Scientists use the data set to identify clusters of particles. They analyze the clusters to generate spherical averages. From these averages, they create .txt and binary HDF files.

Each stage brings the scientists closer to what is in the physical world.

The particles are coordinates on a 3-D grid (512 x 512 x 512) with more grids nested inside. Higher resolution is achieved by storing the position of the particles within the grid.

The purpose of the stages is to go from raw data to data more targeted to a science question. There is a need to identify galaxies or halos on the 3-D grid in order to know where to look in the universe. The 3-D representation is tied to observed quantities, to match analysis.

The third stage is Analysis. This stage gets closer to what is found in publications. Once halos are identified, scientist can look at the image and ask questions about its magnitude is. At this stage scientists are tying the data produced to science questions. The purpose is to tie halos and images together to come up with a number to feed into another plot.

The fourth and final stage is Publication. Analysis creates the images that go with the publication. Any tabular data that goes in to the publication comes out of the Analysis stage.

Once the paper is published, the scientist feels that he is handing off the data set.

3.2 – The Data Table

Data Table Categories:

Data Stage	Output	Typical File Size	Format	Other / Notes
Primary Data				
Simulation	Conditions of the universe at a particular time.	1 TB average Approximately 520 files are generated for each stage (28 stages), ranging from 30-70 GB per file.	HDF5 binary	Tools: custom simulation codes such as INITS, Enzo
Data reduction	Spherical averages of clusters.	2-10 GB per snapshot 20-100 files per snapshot	.txt, HDF5 binary	Tools. HOP halo finder, self-written projection tools, self-written data profiling tools
Analysis	Images of the universe, plotted on a grid.	10-100 MB Approximately 100 data files per snapshot.	images, .txt, HDF5 binary	Tools: MATLAB, IDL, Python, Supermongo. It can be hard to track back from the plots to the files used to generate the plot.
Publication	Final publication with tabular data.			

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.

3.3. - Target data for sharing

The scientist will share data with immediate collaborators immediately after data has been generated, after the data has been normalized and corrected for errors, after the data has been processed for analysis and after the data has been analyzed.

The scientist will share data with others in the field immediately before publication.

The scientist will share data with anyone immediately after the findings derived from the data have been published.

3.4 - Value of the data

The scientist believes that the data from the Simulation and Data Reduction stages have value to other astrophysicists.

Data from the Analysis and Publication stages are easier to interpret and therefore have value to science educators.

The scientist feels that other scientists are interested in the data, but cant' run it at the scale that his group can because his group is a consumer of the supercomputer. He feels that there are lots of people who can ask new questions with his group's simulations.

3.5 - Contextual narrative

This data set is static.

Section 4 - Intellectual property context and information

4.1 - Data owner(s)

Authors of the publication:

Eric J. Hallman
Brian W. O'Shea
Jack O. Burns
Michael L. Norman
Robert Harkness
Rick Wagner

4.2 - Stakeholders

NSF and NASA were the funders. It is unlikely that they need to be consulted regarding the data's release and use because both funders encourage data sharing and data management.

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

None stated.

4.4 - Attribution

Citation.

It is a high priority for this scientist to have the ability to cite this data set in his publications.

It is a high priority for this scientist to require that others cite this data set if they were to use it in their research.

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

5.1 - Overview of data organization and description (metadata)

The data set is organized by timestamps which represent the age of the universe.

A person with some expertise in this field should be able to interpret the data generated at the Analysis or Publication stages. More expertise is required to interpret data at the Simulation and Data Reduction stages.

The scientists would like to publish a list of tools to help users analyze the data.

5.2 - Formal standards used

Data model. Timestamp, snapshot. Every time a new analysis occurs, a new snapshot is generated that becomes part of the data model.

The International Virtual Observatory Alliance (IVOA) is the place to look first for metadata standards.

The IVOA has nothing approaching a standard, but works in data models. Computational astrophysics and astronomy are different in terms of focus. There is no overarching physics data standard. The IVOA resource metadata would apply to the whole dataset and would go into the registry. The registry is OAI-PMH compliant with custom interfaces for getting data out. Each term is defined, for example simulation database or archive. Work like this feeds back into IVOA protocols.

There is an IVOA theory interest group that is driving the organization toward descriptive standards.

5.3 - Locally developed standards

5.4 - Crosswalks

5.5 - Documentation of data organization/description

Section 6 - Ingest / Transfer

The complete pipeline to recapture how all tables and graphs were generated needs to be defined. Necessary scripts need to be defined. Bundle up all data, gathering all intermediate files.

Some data are not as well annotated as other data. Defining metadata for individual file types would be a huge step.

There is a file and then there is the metadata somewhere and the two need to be related. It is a high priority for this scientist to be able to relate the file and its metadata. It is a medium priority for this step to be done in an automated way.

Section 7 – Sharing & Access

7.1 - Willingness / Motivations to share

Publications, almost without exception, are in arXiv.

The scientist is willing to submit data to an open access (public) data repository. After publication, he would agree to full access generally all the way back to the raw data.

It would be beneficial to collaborators if the data could be shared with them from a central repository before publication.

Once published the scientist "would love for the whole world to see it."

The field of astrophysics is small and there is not a need to share with others at the same institution.

7.2 - Embargo

If embargo were an option in the repository, the scientist would deposit data at all stages and embargo until publication.

7.3 - Access control

The scientist would store the data before the Publication stage if access could be restricted to immediate collaborators until publication.

7.4 Secondary (Mirror) site

The ability to access the data set at a secondary (mirror) site is not a priority for this scientist for this data set.

Section 8 - Discovery

"I want everyone to find my data easily."

It is certainly a high priority for people in the discipline to have access to the data – people who would want to do something new with it.

For people outside the discipline, it is not as generally interesting.

People in the discipline should be able to get to the data and to validate the results. A graduate student should be able to access it via Google and to reference it.

Section 9 - Tools

For the published paper, a PDF viewer.

An image viewer, text and spreadsheet tools.

Binary format is usually HDF5 with binary data laid on top of it. All semantics are what the researcher came up with. Researchers usually label data, but users need to know what it is. HDF5 will tell what kind of data array it is, but the user won't know much about it unless he or she knows how it is stored.

For the Data Reduction stage, specialized ENZO analysis tools are needed. There is a publically available tool called YT, written in Python, that has a large user base. The intent is to point users to YT, a scriptable tool that allows interaction with a dataset. Users can download raw data and do something with it.

Section 10 – Linking / Interoperability

If a publication is cited, the scientist wants to refer to its record in astronomical data service. If the scientist publishes a 7th or 8th paper based on a data set, he wants to be able to link to it.

The next archive that this scientist would like to build: an image server to browse images online generated by his group's simulations (64Kx64Kx128K). It would be monstrous, but it would viewable online. He would like to link the image to a simulation that has been curated separately.

Section 11 - Measuring Impact

11.1 - Usage statistics

The scientist places a high priority on usage statistics. It is the only metric the scientist has of who's getting the data other than citations to a document.

11.2 - Gathering information about users

The scientist would like to know the number of unique users and the number of downloads. He doesn't feel a need to know who the users are individually.

Citations. If any persistent ID is used, the scientist wants to be able to track it. It is critical to know how many publications cite the data.

Section 12 – Data Management

12.1 - Security / Back-ups

The scientist stores one copy of the data on disk that is accessible and one copy on tape that is archival (not accessible).

Security measures. The scientist and his group do not allow writing to the data, but there are no other security measures. They do not want anyone to accidentally overwrite the data.

12.2 - Secondary storage sites

For this scientist, It is a low priority for a secondary storage site for this data set.

Section 13 - Preservation

13.1 - Duration of preservation

10 years or more but less than 20

In Computational Astrophysics, simulated results tend to lose value over time. The paper will stay relevant, but the actual data will not stay relevant.

13.2 - Data provenance

13.3 - Data audits

For this scientist, it is a high priority to be able to audit this data set to ensure its integrity over time.

13.4 - Version control

This scientist does not want individual files to be tracked or changed. Once a file is written to disk, it is done. Changes are infrequent enough that the scientist would track them on his own.

13.5 - Format migration

It is a low priority for this scientist to migrate data sets to new formats over time.

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.

14.1 - Primary data contact (data author or designate)

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

14.3 - Campus IT contact

14.4 - Other contacts