Data Curation Profile –Theoretical Chemistry

Profile Author	C.Hudson
Author's Institution	Washington University in St Louis
Contact	C. Hudson chudson@wustl.edu
Date of Creation	March 30, 2012
Date of Last Update	
Version of the Tool	1.0
Version of the Content	
Discipline / Sub-Discipline	Physical/Theoretical Chemistry
Sources of Information	• An initial interview conducted on March 15, 2012 by Rob McFarland, Chemistry Subject Librarian. • Profile edited to include notes from faculty member – April 27, 2012
Notes	
URL	
Licensing	

Brief summary of data curation needs

The researcher has an interest in storing and preserving his data, but questions his ability to sufficiently annotate the data for complete understanding.

Overview of the research

Research area focus

The objective of the research program is to provide a route to the prediction of critical nucleation parameters from computer simulations of the material of interest, i.e., the prediction of the supersaturation required to see nucleation in the laboratory.

The conventional analysis of the nucleation problem invokes "thermodynamic" properties that are not measurable; the essence of the research is that equivalent concepts are provided by statistical mechanics. These can be evaluated at any state of interest and hence the probability of seeing nucleation can be deduced.

Intended audiences

Other researchers in the field, individuals studying nucleation, people using statistical mechanics to deduce thermodynamic properties, and others who are interested in the new methodology or who want to test the method would all be interested in the data.

Funding sources

The scientist sees himself as the owner of the data. Funding for this project has come from the NSF and the University. He has not been mandated to have a data management plan or share or preserve the data. His dataset is not bound by any privacy or confidentiality concerns.

Data kinds and stages

Data narrative

Simulations are performed for a system at a state of interest, observing properties from which the probability of seeing a critical nucleus can be deduced. Some effort is made to map the results found into the conventional language used in this field.

The initial configuration stage of the data cycle involves developing a configuration file to determine the correct atom/particle positions and velocities, the "mechanical state" of the system. There are approximately 9 initial configuration files at this stage. The average file size is 12kb and the data is in Ascii symbols/text files.

The final configuration equilibrium stage of the data cycle involves running the computer simulation or moving the atoms/particles around in space/time to determine if the system has gone into a state of equilibrium. There are approximately 9 data files at this stage. The average file size is 12kb and the data is in Ascii symbols/text files.

The third stage is the simulation - observation and data gathering stage which involves taking the equilibrium configuration and running it through a simulation program that observes the physical property he is researching. This is done repetitively, and properties are observed to determine their actual structure.

The data table

Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
Primary Data				
Config	Atom configuration file	9 / 12 KB	Text/Ascii	
Equilibrium	Final configuration file – final particle location/velocity	9 / 12 KB	Text/Ascii	
Simulation/ Data gathering	For a system that is at equilibrium, 19 different properties are recorded. (Ultimately these are averaged over a time interval to assign an equilibrium value to them.	Varies	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 indicates he would be willing to share the data files with anyone immediately after the results are published as long as they were properly annotated (a description of the content/description of what the data represents) and the data formats explained.

Value of the data
The researcher feels the documented process of the research is of greater value than the actual finalized data. The real value lies in configuration files and how they were organized. Another researcher could use the scientist's data to proof the process.

Intellectual property context and information

Data owner(s)
The scientist believes he is the owner of the data.

Stakeholders
The scientist did not feel there are other stakeholders to the data.

Terms of use (conditions for access and (re)use)
All data released would need to include detailed annotations on process and a data "key" to understand data.

Attribution
The scientist was unsure as to whether or not he would like the ability to cite his data in his publications. The scientist felt it was not a priority to require others to cite the data, access the data from a secondary site or restrict access to the dataset.

Organization and description of data (incl. metadata)

Overview of data organization and description (metadata)
The data is currently organized by an individual schema. The metadata or annotations regarding the data is kept in a diary that describes the research steps including the names of the files and formats.

Formal standards used
No formal standards were used in the organization of the data.

Locally developed standards
The researcher uses his own standards and diary to record his data.

Crosswalks
Not discussed.

Documentation of data organization/description
Not discussed in detail.

Ingest / Transfer

The main issues surrounding ingest of this data into a repository involve annotating the data to a sufficient level of understandability. Since much of the data is not understandable without a code book of explanations, the researcher is especially concerned about the explanations.

He would be willing to share his data and currently does when an individual emails him for more information.

Sharing & Access

Willingness / Motivations to share

The researcher is willing to share, but is worried about whether or not he would be able to provide a coherent image of his data.

He is not willing to share his "diary" or code book of explanations in their current form.

Embargo

The researcher indicated he would be willing to share his data after publication.

Access control

The scientist did not feel it was necessary to restrict access to the data.

Secondary (Mirror) site

The researcher felt it was not a priority to have access to a secondary (mirror) site.

Discovery

The researcher indicated that it is not a priority for researchers within or outside his discipline to easily find the dataset. He does not know what level of priority he would need for the public to easily find his data or for the discoverability of the dataset using Internet search engines.

Tools

To generate the data, proprietary code was written in C and the researcher involved the use of a computer. To utilize the data the researcher wrote his own program.

The scientist indicated a high priority for the ability to connect the data set to visualization or analytical tools. It was not a priority to have others have the ability to comment on or annotate the data set.

Linking / Interoperability

The ability to connect his data with publications or other outputs was a high priority for the researcher. He didn't know if the ability to support the use of web services or connect or merge his data with other data sets was a priority.

Measuring Impact

Usage statistics & other identified metrics

The scientist did not feel it was a priority to see usage statistics on how many people have accessed this data and did not know if it was a priority to track data citations since he believed people would do this regardless.

Gathering information about users

The researcher did not feel it was a priority to gather information about the people who have accessed or made use of the data or to track and show user comments on the data.

Data Management

Security / Back-ups

The faculty member currently organizes his research into directories as an active management practice. He does backup his data continuously but does not take any security measures to protect his data.

Secondary storage sites

The scientist is not interested in secondary storage sites for his data. But he does feel it is a high priority to have a secondary storage site in a different geographic location.

Version control

It is a high priority for the faculty member to allow for version control of the data set.

Preservation

Duration of preservation

The researcher feels as though the data will have value for approximately 3 years or more but less than 5 years.

Data provenance

All parts of the data are equally important to manage and maintain over time. Documentation of any and all changes that were made to the dataset over time was not a priority for the faculty member.

Data audits

The ability to audit the data set was not a priority for the scientist.

Format migration

The ability to migrate data sets into new formats over time was not a priority for the faculty member.