Data Curation Profile – Atmospheric Modeling

Profile Author	M. Cragin
Profile Author	M. Kogan
Profile Author	W.A. Collie
Institution Name	Illinois
Contact	M. Cragin (Cragin@illinois.edu)
Date of Creation	August 14, 2009
Date of Last Update	July 12, 2010
Version	n/a
Discipline / Sub-Discipline	modeling
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 may be 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, approximately an hour long. ▪ A follow-up to the first interview, approximately an hour long. ▪ A second interview with the scientist, approximately an hour long ▪ A questionnaire completed by the scientist as a part of the second interview. ▪ Review of final draft of this document by the scientist
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 Notes	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	The scientist's review of this document has resulted in an Editorial Note, summarizing new perspective based on further consideration of the data for this project.
URL	http://www.datacurationprofiles.org

http://dx.doi.org/10.5703/1288284314997

Brief summary of data curation needs

The scientist expressed an interest in 4-6 month embargo after the publication to allow time for exploration and further analysis.

The ability to connect the data to visualization tools was indicated to be a high priority by the scientist. The data format identified for sharing – Vis5D data – is associated with the open source visual rendering Vis5D software. [SEE the Editorial Note following the Data Narrative] This kind of data could only be opened and worked with in Vis5D software, so the ability to connect the data set to this visualization and analysis tool is essential for others' ability to reuse the data. The scientist also indicated that other web tools and APIs could be useful for the reuse of the data.

The ability for researchers within the scientist's discipline to easily find his dataset was indicated as a high priority.

Editorial Note (Update as of 7/12/2010): The scientist has noted that he now has concerns about the "share-ability" of his data if it is, in fact, deposited in the Vis5D format that is currently used. He states, "For what it is worth, I am moving away from this idea, after seeing its significant limitations: Vis5D works only on Linux and Mac computers, not Windows. There are other (arguably better) choices, e.g. releasing original WRF output (in NetCDF format), reduced/limited WRF output (also in NetCDF), etc. … "The idea was to release a data format that was readily processed and usable in visualization tools, rather than the more "raw" form of the original data." Under consideration is NCAR's Vapor software; www.vapor.ucar.edu).

Overview of the research

The researcher works in the field of Atmospheric Science, for which there are (at least) two approaches to research within Atmospheric Science – field research and computational modeling. This scientist uses computational modeling to produce simulations of weather phenomena. The goal of his work is the development of predictive models for severe weather activity.

Research area focus
This research project centers on modeling of severe weather and the interaction of storms. The research investigates the impact of the relative location of storms, in combination with other parameters (i.e. variables) on subsequent weather development. Modeling methods require data for input to represent the variables and context in order to produce simulations; these input data are generally acquired from external sources. There are two possible types of input data for the weather simulation: observational data collected during real weather occurrences, and "idealized" data are derived computationally from observational data through the application of rule-based algorithms. This computation produces systematic changes (and data reductions) that result in a simplified "typical" weather profile (temperature, pressure, moisture, wind, etc. at each height) for a particular weather phenomenon. Both real and idealized data can be obtained from the National Climatic Data Center website (http://www.ncdc.noaa.gov/oa/ncdc.html). This scientist usually uses "idealized" data as input for his simulations.

Intended audiences
Possible audiences for these data include atmospheric scientists interested in using weather simulations for educational purposes; numerical modelers wanting to reproduce this scientist's results; and scientists who use data mining techniques to elucidate larger weather patterns through analysis of large aggregated data sets.

Funding sources
National Science Foundation (NSF)

Data kinds and stages

Data narrative

One part of the data consists of the various parameters, settings, and version of the Weather Research and Forecasting model (http://www.wrf-model.org/index.php), which is a community model for weather forecasting and currently is a community standard. Another major component of the data is the output of the model, which goes through multiple transformation stages.

In the initial step, the typical weather profile of temperature, pressure, moisture and such (called "the sounding") gets combined with the "namelist" file, which contains the model options representing the physical processes in the atmosphere and the storms' relative locations. The "sounding" and the "namelist" are compiled together by the pre-processing module of Weather Research and Forecasting Model called "ideal.exe" (see *Initiation for Ideal Cases* in http://www.mmm.ucar.edu/wrf/users/docs/user_guide_V3/users_guide_chap4.htm). This compilation produces ready-to-use "idealized" input that now can be used to initiate the model.

The output of the model is three-dimensional floating point data that is considered the raw data for this research area (i.e. sub-discipline). This data is output from the model at regular intervals of time. The output at each time step is saved as a separate file, since appending all the output to one file makes the file too large. The raw output data is uniform in its horizontal dimension, but not in the vertical one due to terrain differences. Thus, the raw data has to be made uniform all around by interpolating it onto the regular Cartesian grid. This interpolation is done by putting the raw data through a post-processing tool called "Read/Interpolate/Plot" or "RIP." (http://www.mmm.ucar.edu/wrf/OnLineTutorial/Graphics/RIP4/index.html).

While making data strictly Cartesian, RIP also simultaneously produces the basic radar-like plots from the data. The scientist considers this to be the first generation of images, since other images of various quality and formats will be generated from these basic radar-like plots. The various images, in turn, will be turned into animations, as well as put into PowerPoint presentations – an important method for sharing data in this research community. The interpolated Cartesian data, meanwhile, gets further reduced into statistical summary files that summarize the changes of various weather characteristics over time. The statistical summary files together with all the visual outputs (plots, images, animations) get combined into a web page for the simulation. The statistical summaries are also further condensed into a spreadsheet that compiles the analyses across multiple simulations.

See table on the following page

The categories in the "data stages" 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
"Raw"	Floating point 3-D data	181 files X 100-200 Mb each	NetCDF	A file is spat out at each time step
"Interpolated"	Cartesian interpolated data	Smaller than raw: < 100 Mb each	Raw binary .dat	181 files
"Intermediate"	Compressed binary ASCII form of raw data	10-100 Mb	Vis5D .v5d format	1 file per dataset
"Analyzed"/"Statistics"	Statistical summary of data	trivial	.txt	1 file/simulation
Overview	Web pages with graphics and stats	Around 140 Mb	.html	
Final	Statistics across multiple simulations	small	Excel spreadsheet	
Ancillary Data				
Text	n/a	5Kb	Fortran	"namelist" file
Text	n/a	8Kb	(unspecified)	"sounding" file
Image	1.Images 2. Animations	1. Less than 1 Mb 2. 5-6 Mb	1.NCAR graphics metacode, Computer Graphics Metafile, Adobe Illustrator files .gif 2. animated gif/ QuickTime .mov	

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 data most suitable for ingest is the intermediary data form in Vis5D format. This form of the data is less voluminous than the raw or interpolated data, but it retains most of the information necessary for visualization and analysis. Since Vis5D is an open source multi-functional data visualization and analysis program, the data in the corresponding format is quite useful for sharing, and thus would be the target data for ingest.

Value of the data

This scientist reports that the most informationally valuable part of the data he produces is the "Intermediate" model output data, which is "a compressed binary form (typically 4:1 compression, 1 integer per floating point number). [A]ll desired data times [are] included in this one file, ready for use by the Vis5D display program. Vis5D format includes all (desired, could be subset) data times, with floating point data reduced to 1 byte for each number. [With this format], the (linux/mac) user can

immediately visualize and animate the fields from this compressed form. The file typically contains only a few fields from the original WRF output (e.g. temperature, wind, reflectivity – not the huge list of fields routinely output from WRF)."

Because it is easy to replicate, the data from this type of project has limited value for long-term preservation; it is, however, valuable for integration with other datasets for the purposes of data mining.

In addition, there are plans to do more longitudinal modeling and the scientist believes that data would be more valuable because it is much more difficult to replicate.

Various aspects of the current data have re-use value for different audiences:

- General atmospheric science community could use the radar plots, animations, and reduced data behind them as educational aides.
- Other computational modelers would use the "idealized" weather input, model parameters, version and settings to reproduce the results.
- The data mining analysts would need the raw output of his simulations, along with "the input parameters, in order to develop rules and other associations between the specification of the problem (e.g. parameters for each run), and the resulting output (mined by them)," i.e. these new users.
- So would animators interested in producing a high-quality animation. Generally speaking, the scientist believes some reduced version of the raw data would be the most useful to others.

Contextual narrative
The data set is static since it results from a finite simulation that aims to model a weather phenomenon of a specific finite duration. The duration of the simulation and frequency of the data output from the model are determining factors in the size of the data set. With longer simulations that maintain high frequency of data output, the raw data easily gets to hundreds of gigabytes.

The outputs of the simulations include the NetCDF raw gridded floating point data, binary (.dat) interpolated Cartesian data, intermediate Vis5D compressed binary data, text file statistical summaries for each simulation, and statistical summary across the simulations in an Excel spreadsheet. The author also produces multiple types of images: initial NCAR graphics metacode plots, more advanced images in the standard Computer Graphics Metafile (CGM) format, and publishable quality images produced by Adobe Illustrator. The images are often combined into animations, which the scientist saves as animated .gif or QuickTime movies. All the statistical and most of the visual information produced about each simulation is combined and made available though a web page.

Intellectual property context and information

Data owner(s)
This scientist considers his status as that of "holding" the data, and maintaining it for his own use. However, eventually the public would have full access to it and own it, since it was funded by public monies through NFS.

Stakeholders

- NSF as a funding agency
- The National Center for Supercomputing Applications (NCSA) as providing a service supporting his data
- His collaborators

Terms of use (conditions for access and (re)use)
Outside of immediate collaborators, the data could be made fully available to others as long as the project is far enough along in terms of the publication timeline. The scientist wants to be sure he is able retain the credit for publishing based on the data.

Attribution
The scientist indicated that the ability to cite the dataset in his publications is a high priority. The scientist is also concerned with other people citing his data and giving proper attribution, which according to him happens only every other time.

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

Overview of data organization and description
Raw data is stored in the NetCDF – Network Common Data format (http://www.unidata.ucar.edu/software/netcdf). This format includes a header describing the layout of the file, as well as the names of the attributes.

As noted above, the most shareable form of the data has been identified as the “intermediate”, or compressed data form, stored in the open source Vis5D format (http://www.ssec.wisc.edu/~billh/vis5d.html), which the scientist believes is “a self-described data format.” This format supports the organization of five-dimensional (spatial 3D, time, physical parameters) matrix data by including the names for all the attributes.

This scientist does use a systematic naming convention to organize the output of his model runs. The wrfout* files are the raw NetCDF output of the model, which are saved out at regular intervals of time. The star refers to the time elapsed from the beginning of the simulation: the file ending with ‘:00’ is the first output of the model, followed by one ending in ‘:05’ if the time step selected is 5 minutes. The interpolated Cartesian files are named wrfout*.dat, appending ‘.dat’ to the full names of the raw output files. The web data (images, animations, text, html), as well as other files (NCAR Graphics metacode, suitable for conversion to CGM and then to Adobe Illustrator) are all stored in "run_archive.tar". All these files are stored in one directory representing a particular model run, named WRF-‘run number’. There is a particular file structure imposed by the scientist for the management and storage of these data, both locally and for the data sets on the TeraGrid.

Formal standards used
The several data formats that are used (NetCDF, CGM, Vis5D) have been described above. No formal metadata standards, ontologies, or controlled vocabularies have been applied to this data.

Locally developed standards
None.

Crosswalks
Not discussed.

Documentation of data organization/description
The scientist did not talk much about a system for documenting these materials, though sometimes a logbook is kept in Excel. “Data are archived for long-term storage on NCSA’s mass storage system “Unitree.” In general, the goal is to store all relevant files that would be needed to reproduce a simulation; however, this is somewhat incomplete in that some aspects of the computational system used are not recorded. The data are, though, complete in the sense that the input namelists & sounding, and standard NetCDF output files are all archived for each run.”

As noted above, initial parameters for simulation "runs" are loaded into two text files, which act to start the computation. While simulation functions are set for individual experiments, processing activities for this area of research appear to be fairly common, and the set of parameters (and other "in-put data) are documented in the methods sections of published papers.

Ingest

Upon review of this document, the scientist indicates an automated process for submission to the repository is a high priority for this kind of data, particularly as ingest might become a component of the java workflow system, and the upstream processes need to be part of what is captured and deposited. The ingest process must be "as painless and comprehensive as possible."

Access

Willingness / Motivations to share
Researcher is generally open to sharing with various groups, but does not have much experience with sharing beyond the immediate collaborators. Most of his sharing has occurred with immediate collaborators at the same location.

The scientist indicated that he would be willing to share the raw and corrected stages of the data only with his immediate collaborators. After the data has been possessed for analysis, he would be open to sharing with others in his research institution. The same applies to the analyzed data. Immediately before publication and right after data have been published, the researcher sees it appropriate to share the data with his professional societies. He is willing to share the data with general public only after an embargo of about six months.

The period of embargo is necessary to ensure that the scientist gets an opportunity to explore and publish on all the aspects of the data that interest him. In the meantime, the scientist envisions different groups having a different degree of access to the data in the repository. For example, some groups should have restricted access, while others (like NSF and reviewers) should have the data fully available.

Embargo
Six months of embargo is seen as adequate by the scientist for the most informationally valuable intermediate Vis5D data, since it is sufficiently removed from the very raw data.

Access control
The ability to restrict access was identified as a medium priority.

Secondary (Mirror) site
The ability to access the data set at a secondary (mirror) site is low priority.

Discovery

This scientist is interested in the availability of different methods for locating these data in a repository, including browsing capability and an index. He expressed concern for potential users who might not be familiar with disciplinary pathways. He suggested that university departments have links from their websites, and references to the repository to increase data's visibility. Another potential option would be to link to existing dissemination pathways, such as the American Meteorological Society's website. The ability for people to discover his datasets using internet search engines such

as Google is a medium priority, and while he thought that it would be useful to have additional search capability, he did not specify what this might be.

Tools

While certain data analysts make use of his data for the purposes of data mining, it is not clear that data mining tools would be of any use to the majority of data users (including the scientist). The researcher indicated that connecting data to visualization and analytical tools would be a high priority for this kind of data. No analytical tools are needed for opening the most informationally valuable Vis5D data. [Update: As stated in the Editor's Note (page 2), there is consideration now for moving to a different file format that would support even greater access to the data.]

Interoperability

The scientist did not mention the need to be able to interoperate this data with other data sets or to include this data in federated searching to make it more discoverable.

The scientists did indicate that availability of web service APIs is a high priority for his data. He also specified ability to connect data to tools to be a high priority.

Measuring impact

Usage Statistics
The scientist indicated that measuring usage statistics would be a low priority for his data.

Gathering information about users
The scientist did not specifically discuss gathering information about users.

Data management

Security/Back-ups
Since the data is a result of a simulation and it doesn't carry any privacy/confidentiality/security risks, the scientist is not interested in any advanced security measures. He also believes that encryption would slow down the transfer of already large datasets, so it is more of a liability than an asset for this data.

Some of the data is stored on a local PC, and though it has been turned off for the time being, there is a script that is set to run weekly to back-up the data to another disc drive. The bulk of the data is stored at the TeraGrid at NSCA, so in the past the scientist has not had to worry about the maintenance and back-up of data stored through that service. However, NCSA "now requires continued and select requests (e.g. each year, through TeraGrid) for mass storage resources; there is no open-ended repository for the data per se."

Secondary storage sites
A secondary storage site is a low priority for the scientist, partly because he can easily regenerate a lot of his data, and in part because of NCSA's TeraGrid service.

Preservation

Duration of preservation
In the interview scientist indicated that he tries to keep the raw model output data indefinitely, since he can produce all the other data stages from it. While currently it is easy to reproduce, the researcher intends to focus more on longitudinal models, outputs of which would be more difficult to regenerate. Thus his general practice is to keep the raw data indefinitely. The data interpolated to the Cartesian grid is usually kept for the duration of the project. For the intermediate compressed Vis5D data, the scientist indicated that the data ought to be maintained in a repository for about 15 years, especially if it is cited in published papers.

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

Data audits
The scientist indicated the ability to audit the dataset to be a medium priority for his data.

Version control
Version control is not applicable to this dataset.

Format migration
The scientist indicated the ability to migrate the dataset into new formats over time to be a medium priority for his data. Most formats used by the researchers are freeware and open source formats.

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