Data Curation Profile – Aerospace Engineering / Chemical Kinetics

Profile Author	Nabil Kashyap
Author's Institution	University of Michigan
Contact	bil@umich.edu
Researcher(s) Interviewed	[name withheld], Doctoral Candidate, Aerospace Engineering
Researcher's Institution	University of Michigan
Date of Creation	March 25, 2012
Date of Last Update	March 25, 2012
Version of the Tool	1.0
Version of the Content	1.0
Discipline / Sub-Discipline	Aerospace Engineering / High Temperature & High Pressure Chemical Kinetics
Sources of Information	• An initial interview conducted on March 25, 2012. • A sample of the profiled data as well as literature authored by the interviewee.
Notes	Instead of administering a worksheet, the interview was conducted orally in its entirety from the worksheet questions.
URL	http://datacurationprofiles.org
Licensing	

Section 1 - Brief summary of data curation needs

In the process of running an experiment, this researcher generates a variety of files from text to video, from open source formats to proprietary files specific to the equipment involved. His entire output for his dissertation research amounts to approximately several gigabytes of data. Outside of datasets downloadable as supplemental information maintained by discipline specific academic journals in which he has published, there are no formal venues for sharing data. Protocols for data management are highly researcher-specific in his lab, though data on lab servers is backed up regularly. There are no specific standards for either data management or sharing within the institution or even between labs.

http://dx.doi.org/10.5703/1288284314989

Section 2 - Overview of the research

2.1 - Research area focus

This researcher's recent work focuses on the ignition delay of fuel-air mixtures under high temperature and pressure conditions using a rapid compression facility, which consists of a tube 8.2m long through which a free piston is applied using highly compressed air. By sampling and imaging gases in the short interval before and during ignition, these experiments shed light on the formation of reactants -- including greenhouse gases, particulates, pollutants and other products of combustion. He studies both global reactivity and individual reaction pathways of particular species of reactants.

2.2 - Intended audiences

Data from his lab are typically of interest to a very select range of disciplines, primarily for chemical kineticists; however, results might be of interest to atmospheric chemist and possibly engineers designing combustion engines, for instance, in the automotive or jet industries. Reactants and pathways by which reactants form are significant to understanding changes in the atmosphere and to improving the design of combustion engines. On the other hand, this researcher was unsure how data from these particular experiments might be used in other contexts.

2.3 - Funding sources

The research is primarily funded by the Department of Energy, the Graham Environmental Sustainability Institute and discretionary funds from the Department of Mechanical Engineering. None of these sources require a data management plan of any kind.

Section 3 - Data kinds and stages

3.1 - Data narrative

In each experiment, a piston drives the volume of a tube 2.7m long into a much smaller test volume, compressing a fuel-air mixture into a 100-200cc and igniting the mixture in the process. This researcher's experiments fall roughly into two categories: those which capture video of the 200ms-long process and those which take physical samples of the mixture at different stages during the reaction process. Those samples are then separated using four chromatography machines.

The resulting data is captured in several formats: raw data files from the pressure transducers in the compressed tube and from the chromatographs as well as proprietary digital video. At the peak of activity, he might run four experiments a day -- however, periods during which he actually runs experiments may be broken up by months of research attempting to formulate hypotheses to test.

The data is then analyzed, which involves generating proprietary files for processing software and convenient printable formats for manually examining the data, for example Excel spreadsheets or Portable Document Format files. Pressure trace graphs and chromatographs are the focus of analysis, while video from the experiment is used primarily for verification that the experiment ran correctly. Finally, the data are readied for presentation through generating stills from the video files, compiling plots through another proprietary software and sometimes combining relevant stills and graphs together using Photoshop. Analyses are also written up using LaTeX and Word. The researcher did not distinguish between preparing data for conference presentations as opposed to publication.

3.2 – The data table

Data Stage	Output	Files per experiment / Typical Size	Format	Other / Notes
Primary Data				
Raw	Pressure trace	1 / 5mb	.dat	Generated from pressure sensors using Labview
	Concentrations of constituent reactants	4 / 100kb	.dat	Generated from chromatographs
	Video of ignition delay	1 / 50mb	.cin	CCD camera firing 30,000/s (Phantom high-speed camera hardware/software)
Analyzed	Pressure trace data	1-3 / 100kb	.mat	Data analyzed using Matlab
	Pressure trace data	1-3 / 100kb	.xlsx / .pdf	For convenience, also analyzed using Excel (Office Open) and Adobe Portable File Format
	Chromatograms	4 / 100kb	.prm	Chromatograms interpreted for Clarity software
Finalized	Arrhenius plots / concentration plots	15 - 20 / 1mb (per article or pres.)	.jpg / .pdf / .eps	Graphs generated using Origin software
	Video stills	1-5 / 1-5mb (per article or pres.)	.jpg	Exported images using Phantom software
	Photoshop composites	1-2 / 1-10mb (per article or pres.)	.jpg / .psd	Composites generated using Photoshop merging graphs and video stills

3.3. - Target data for sharing

The researcher suggested a willingness to share most data but expressed reservations about what would be actually useful and about a discipline wide concerns regarding varying degrees of academic dishonesty. Raw data files are most likely to be useful for analysis but potential user-communities of raw data would be severely limited to a few specialists who could interpret the data taking into account the specifics of the lab's particular equipment and calibration.

3.4 - Value of the data

While in practice the data, the researcher expressed doubts that the data would be of immediate use to anyone other than a handful of experts in the field, there is the potential for the data to be of interest to atmospheric chemists and designers concerned with high-temperature / -pressure combustion engines, including automotive and aerospace applications.

3.5 - Contextual narrative

While it was clear that the data supporting pressure traces and Arrhenius plots were the focus of his research, in order to be fully understood, a researcher would need a full complement of data -- including imaging, chromatographs and some understanding of the facilities. According to him, the technical specifications of the lab and how it was calibrated are critical to effective interpretation: "Where the data come from is as important as what the data are -- there are assumptions in the system that change the interpretation of the data."

Section 4 - Intellectual property context and information

4.1 - Data owner(s)

According to this researcher, rights to the data belong to the University of Michigan Board of Regents; however, he did not see this as impeding collaboration or sharing among invested parties.

4.2 - Stakeholders

The researcher expressed a potential need to consult with his faculty advisor for advice and permission with regards to releasing data.

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

This researcher said that, in order to be in keeping with his discipline's practices, he would be interested in sharing more polished forms of data as watermarked PDFs in order to avoid possible issues of plagiarism. On the other hand, he was more concerned that data be useful -- that it undergo verification and some vetting as to the certainty with which concrete conclusions might be drawn before data were released.

4.4 - Attribution

Attribution is a high priority. According to him, the culture of the field is fairly protective of their data and that his views vary even from those with whom he shares his lab. He suggested that his colleagues might be less likely to share data.

Section 5 - Organization and description of data

5.1 - Overview of data organization and description (metadata)

The data are currently organized and described in a manner highly contingent upon the individual researcher and culture of the particular lab. Description, organization and analysis are conducted by the primary researcher in conjunction with the researcher's advisor. Those practices appear to be influenced by labmates who often assist in performing each others' experiments but may not otherwise be formally working with each other. For example, he indicated that he generated more files during the analysis stage than his colleagues because of preference, but that he may be less thorough in providing metadata like context clear column headings. This indicates that labmates are aware of and take into account each others' data management practices. In his view, metadata is context specific -- depending on the intended audience, he adjusts the degree of description that accompany the data.

5.2 - Formal standards used

The researcher did not indicate any discipline specific formal metadata standards were either required or employed.

5.3 - Locally developed standards

The researcher did not indicate any locally developed metadata standards were either required or employed. Instead, metadata appears to be up to the discretion of the researcher as part of one's personal research habits.

5.4 Crosswalks

Not discussed

5.5 - Documentation of data organization/description

Not discussed

Section 6 - Ingest / Transfer

While the researcher has not transferred his data to a repository, he indicated that before doing so he would want to amend appropriate metadata such as data column headings and chart legends. He was particularly concerned about two issues: 1) That enough metadata was packaged with his files to describe the specifications of the lab and 2) that he be able to determine which data was appropriate to preserve based on perceived usefulness and the certainty with which they might support particular claims. The data as it exists on the server and as it might be informally traded among similar departments across institutions would not be useful outside a small community without significant preparation prior to ingest.

Section 7 – Sharing & Access

7.1 - Willingness / Motivations to share

He was enthusiastic about the ideas of interdisciplinary reuse and of long-term preservation but did not see either discussed in his field in a serious way. For example, depositing data into repository is not an encourage practice in his discipline.

7.2 - Embargo

The researcher did not feel the need for any kind of embargo. The most important criterion was that the data be useful, that he have some degree of confidence in them before sharing. This time for analysis could de facto be a kind of embargo.

7.3 - Access control

He personally was not concerned with access control. Public availability was much less a concern than researchers within his discipline possibly engaging in academically dishonest behavior. Ultimately, he did not appear sure how useful the data would be outside those specifically trained to interpret it.

7.4 Secondary (Mirror) site

The need for a mirror site was not a high priority though the usefulness of an offsite server was immediately apparent to him. He related a recent incident in which labs connected with his -- labs with more lax backup procedures than his -- recently suffered massive setbacks due to user error in an unaffiliated department that caused desktops across the network to format their hard drives.

Section 8 - Discovery

Currently his data are only available by either contacting him directly. A portion of his data are available as supplementary information via websites of academic journals in which he has published; these journals are in turn indexed by aggregates such as Web of Knowledge and Science Direct. He rated discovery a high priority, both by those within his discipline and those outside the discipline -- in chemical kinetics, no such mechanisms exist. On the other hand, he expressed skepticism regarding the actual reuse of his data by those outside his field.

Section 9 - Tools

Creating these data involves a variety of hardware and software products, from high-speed cameras and laser triggers, to pressure transducers and advanced imaging software. To analyze the data to the same degree that he is able requires several specialized pieces of software, primarily Phantom, Matlab, Clarity and Origin. On the other hand, the data are accessible through well-documented formats such as Comma Separated Values or Microsoft Office Open standards. Image data is often exported to JPEG but can be exported to any number of formats, including

TIFF and PDF. In terms of use, the ability to visualize the data is a high-priority while the ability to annotate is not entirely relevant to the research.

Section 10 – Linking / Interoperability

The researcher indicated that interoperability was of great interest to him. The ability to link published articles with data is of particular importance and is a standard practice in the major journals of the field. The ability to access the data through APIs seemed like a very useful tool, particularly a sort of dataset search engine. Merging datasets also seemed very promising but because the data is so lab-specific, he pointed out that it would be critical to do so responsibly.

Section 11 - Measuring Impact

11.1 - Usage statistics & other identified metrics

The researcher was very interested in identifying not only how much the data was accessed but also by whom. More specifically, he would like to know what fields of research and what institutions might be using the data.

11.2 - Gathering information about users

Not discussed.

Section 12 – Data Management

12.1 - Security / Backups

Current backup practices are adequate according to the researcher. His lab's RAID array is scheduled to perform automatic backups every week. All data is password protected and only accessible onsite.

12.2 - Secondary storage sites

While there are several servers to which copies are saved, there is no offsite mirror. All drives are within the vicinity. The researcher expressed no concerns about this; however, it is worth noting that these experiments consist of controlled explosions and require a quantity of highly combustible fuel to be on hand. The rapid compression facility and the fuel are located in the same building, sometimes the same room, as the servers.

12.3 - Version control

The researcher was intrigued by the idea of version control but had no processes in place. Formally tracking changes seemed like potentially useful metadata but he did not currently create that kind of data nor could he immediately see how his current software would automatically allow him to do so.

Section 13 - Preservation

13.1 - Duration of preservation

A great deal of research goes into each round of experiments (as opposed to more high-volume research being done in the same lab) and the data is not time-sensitive -- if the data were to be preserved, it should be for an indefinite term.

13.2 - Data provenance

Provenance is one of the most important parts of to preserve of this dataset because the conditions of the lab could inform how the data are interpreted. To extrapolate somewhat,

descriptive information regarding the lab equipment would be helpful, but actually linking the data to this lab, at this particular point in the lab's history, under this advisor, would allow a future researcher to reconstruct initial conditions most accurately.

13.3 - Data audits

The researcher indicated that checking the integrity of the data was a high-priority -- without confidence in their integrity, the data are no longer useful.

13.4 - Format migration

While migration is a priority, the scientist said that he would like to see planning in terms of initial ingest, for example, restricting data to open formats that would require little or no migration.

Section 14 – Personnel

Withheld from the public version of this profile.