Fully-Connected Scientific Metadata

Joe Edelman
Dartmouth College *
joe@orbis-tertius.net

December 12, 2003

Abstract

This paper concerns desirable properties for scientific data in databases. A metadata scheme is introduced that focuses on the *statistical comparability* and *completeness* of data objects, rather than on IP issues, access privileges, provenance, annotations by domain experts, and other topics that have been well-treated.

Notable characteristics of the scheme include (a) a characterization of measurements and recordings that includes physical units, scales, and uncertainties, rather than just numbers, (b) a rigorous accounting of the processes by which data is selected, summarized, and transformed from one scale to another, and (c) a characterization of an entire study as a well-connected object, rather than as a set of individual data files.

1 Introduction

Scientific data is similar to other multimedia in some respects. The following questions are equally relevant, whether asked about blockbuster movies or scientific datasets:

- Who are the authors?
- Is it widely respected?
- May I download it?
- What are keywords or topics of relevance?

*This work was conceived at the fMRI Data Center under the direction of Jack van Horn and Jeff Woodward, and supported by NIH grant X and NSF grant Y.

But there are several questions one might ask about scientific data that would be out of place when asked about other multimedia, such as:

- Is this statistically comparable to my own data?
- What are confounding factors and sources of noise in this data?
- What machines and configurations were involved in the recordings?
- Are all steps of the analysis, from recording to the published figures, accounted for rigorously?

This paper explores the possibilty of automatically answering these latter questions based on a metadata schema specific to methods in the physical sciences.

2 Metadata Implications of the Scientific Process

Sociologists have studied the unique ways that data are used and shared in the scientific process.[1] Building on their work, we present three features of science which impose requirements on metadata: the openness of the scientific process, the need for rigor in analysis and comparison, and the relation of data to the real world phenomena it encodes.

2.1 Open Process

In the film industry, the important thing to share is a set of end products: trailers, movies, shorts, etc. These can be packaged up neatly using existing formats, although arranging a quantity of them in a browsable, searchable database is still challenging.

In contrast, metaänalysis, reinterpretation, collaboration, and peer review in science require the entire sequence of operations involved in producing the final products (statistical determinations; confirmation or rejection of a hypothesis) to be made clear. Furthermore, due to the frequency of reinterpretation and metaänalysis in science, intermediate data objects (recordings, etc.), when considered with good information about the methods and procedures involved in their production, may often have more lasting value than the figures and conclusions scientists draw from them at the time of publication.

Thus it is important for metadata in scientific databases to answer questions about figures like "how was this made?" An even broader and more interesting question might be, "if I had the appropriate software, could I

rerun the analysis here locally, starting from raw data, and obtain the same figure?" To answer questions like these requires metadata fields both for (a) relating each data object to its antecedents in the analysis-recording chain (see section 3.1), and (b) describing the events and participants involved in any recording session, so that alternate interpretations of recordings can be made (see section 3.3).

2.2 Rigorous Math

In the audio and video editing industries there are advanced numerical algorithms that use many of the same techniques used in processing scientific data: smoothing, components analysis, etc. There should, however, be a difference in how these algorithms are selected and combined: In the world of art, the aim is make something that captures the artist's intent or the delights the viewer. With science, the aim is to rigorously support a claim.

New techniques for data manipulation and statistical analysis are invented daily. With significant exceptions, these are well-founded. This is because scientists, in their initial design and documentation for a new technique, consider the statistically relevant parameters (quantity measured, distribution, scale, uncertainty, significance, etc.) of their data fed to the new technique, and use that information together with the structure of the technique to evaluate the meaningfulness of the results. Statisticians excel at this kind of reasoning about which techniques are meaningful in which situations.

As the sources of our data become more heterogeneous, though, and as opportunities for data input multiply, this process, which amounts to the verification of *statistical comparability*, becomes an unbearable burden on the scientist. To evaluate the statistical comparability of two datasets requires good information about the scales and uncertainties of data files (see section 3.2), as well as information about the distributional consequences of previous algorithms through which the data has been run (see section 3.1). To aid the scientists in searching for data for use in a metaänalysis, we need to couple that metadata with methods for relating the inputs requirements of statistical methods with the properties of the data files available (see section 3.2.

2.3 Extrinsicality

We say that movies are *about* an event, that documentaries are *on* an event, subject, or object, but a scientific recording is *of* some particular object or

phenomenon in the real world. When our software can figure out exactly what a given recording is *of*, this can provide a tremendous advantage in searching for relevant, comparable data.

This third notion is much better accounted for in existing systems than the two treated previously. For instance: BioImage[2,3] has an elaborate ontology for specifying the chemicals, strains, cell types, etc. which a cell microscopy recording is *of*. The NeuroNames[4] project has a detailed breakdown of brain anatomy and function, all the way down to individual neurons, for specifying what a brain scan or electrode recording is *of*.

The representation of these notions in metadata, however, is vastly different in each domain. For cross-domain searching and for radical reinterpretation (e.g., for claiming that the reported activation in an fMRI scan was due to blood-flow/vasculature considerations) it would help to have an overarching framework for describing *all* the events and participants involved in any recording session (see section 3.3).

3 Fully-Connected Scientific Metadata

Consider the requirements for statical comparability. For two datasets to be comparable, they must be

- recordings of comparable *phenomena*
- measured on comparable *scales*
- which have undergone comparable *pre-processing*

These three correspond roughly to domains within our metadata scheme.[1] The resulting whole provides support for evaluating, not only the statistical comparability of data objects, but also their completeness, rigor, possible reinterpretation, and utility for metaänalysis.

Section 3.1 discusses a simple model for encoding the computational history, or *pre-processing*, of data, through the linking of data objects and processes. Section 3.2 presents a characterization of *scales*, via dimensioned linear algebra. Finally, in section 3.3 we advance some very general ideas about packaging descriptions of recorded *phenomena* so that they can be shared across disciplines.

[1] An open source proof-of-concept implementation of the metadata framework described in this paper exists as an ontology in OWL and as a *Protégé* knowledgebase. It can be downloaded from the web.[5]

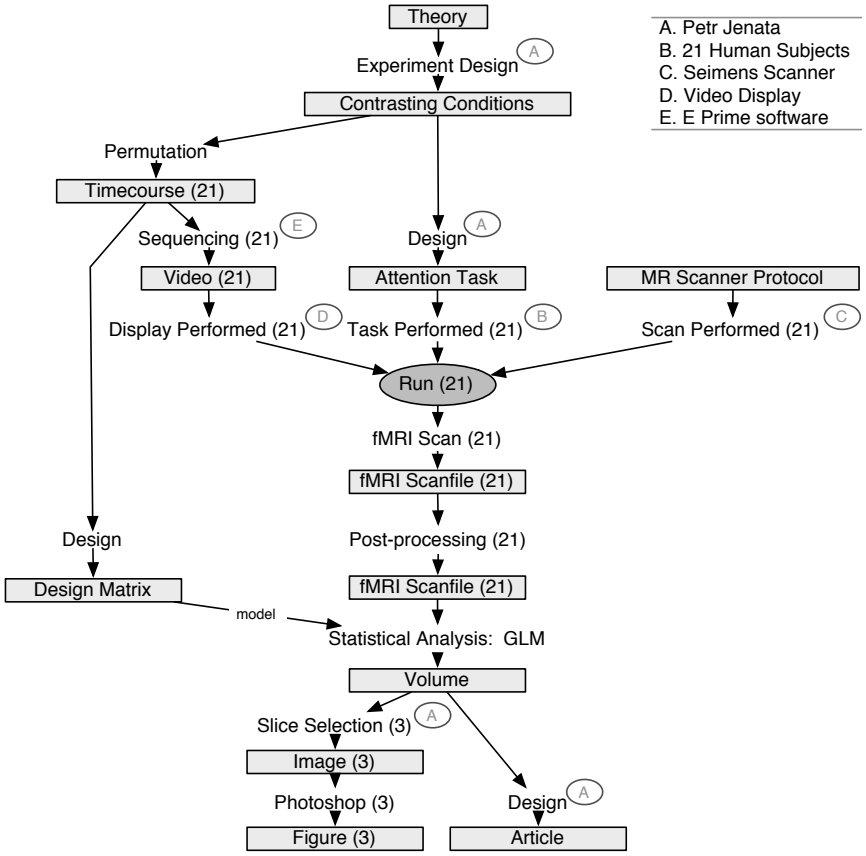

Figure 1: Illustration of the connections among data objects and processes in a typical fMRI study. In databases based on our metadata scheme it is possible to provide a graph view like this to the end-user. Several features are worth noting: (a) there is a strict alternation between data objects (in boxes) and processes; (b) nearly identical data objects and processes have been coalesced (e.g., the 21 scanfiles); (c) the use of photoshop in figure preparation is crystal clear; (d) the use of a model-based statistical technique is also made clear.

3.1 Data Objects and Processes

The problem of determining whether two data objects have undergone comparable processing is addressed in two steps: first, by defining a representation for the *computational history* of each data object by linking it to it's *antecedents*; then, by then using this computational history to decide a notion of comparability.

3.1.1 Nomenclature

A *file*, or more generally a *data object* or *product*, results from the *execution* of some *process* with some other files or data objects as *input*.

A *design process* involves subjective *human input*.

An *analysis process* is automated, and may be either *model-agnostic*, in which case its input it only the data and a small number of parameters, or *model-based*, in which case it has a *designed model* as input as well, and the output consists of correlations between the data and this model. As black boxes, model-agnostic analysis processes may be characterized in information theoretic terms, and they generally provide some kind of selection of, compression of, elaboration upon, or other computation upon the input.

A *recording process* involves a set of external *participants* including *instrumentation* (which is subject to a *configuration*) and the objects of study (which are subject to a *treatment*). More is said about recording processes in section 3.3.

3.1.2 The Flow of Events

Our specification of computational history imposes a partial order on all components of a study (figure 1), and leads to an *experimental overview graph*. This graph is directed, acyclic, and places those things that one begins an experiment with at the top, and those things one has when one finishes at the bottom. Inputs are represented as lines from data objects to processes, and outputs from process to lines. Physical and human participants, such as recording devices, researchers, and the objects of study, appear as floating labels. Where there would be several identical nodes (i.e. when several recordings are made of different objects with the same treatment, and they are analyzed identically) these are coalesced.

Although alternate structures are easy to arrange, this graph tends to emphasize the natural structure of most scientific experiments, which proceed from a theory to an experimental design to a set of *treatments* and

configurations for the objects and instruments of study, to a set of recordings, to analyses, to figures and journal articles. The graph is narrow at the top, consisting of ideas and experimental designs, and at the bottom, with only a small number of final products. In the middle, where the data is, it is wide.

3.1.3 Capturing the Information

While a process is running, there is information available to the operating system about its nature and about which files are used as input. When the process completes, however, this information is seldom stored anywhere or attached to the files involved.

This information loss is a big problem for scientific data sharing. Scientists end up with a large collection of files but none of the connective, process-level information necessary to make a complete story about an experiment. As this information is often necessary for other scientists to understand, reproduce, or reuse data from the experiment, much of the potential utility of the data objects to other scientists has been lost.

That important information is lost is widely recognized, and many "image annotation" efforts have focused on returning this kind of information to images in scientific databases,[2,6] while other, newer efforts in particular instrumental domains like microscopy[7] have developed platforms that avoid the loss of information in the first place.

It is likely that this approach will be generalized by the combination of (1) a domain-independent metadata model capable of absorbing this information with (2) a desktop, database, and data workbench-like platform which tracks and connects together the various data objects and processes used in an experiment as the user operates upon them. Both the Semantic Web initiative[8] and the European eScience infrastructure[9] adopt this vision, and there are already some software packages that hint at its realization.[7,10–12]

3.1.4 Comparability

Here we discuss only aspects of comparability related to what we have called *pre-processing*, or *process comparability*. Specifically we leave discussion of whether the various processing done on a data object has cast it into an incompatible *scale* or *statistical distribution* to section 3.2, and we leave discussion of the different kinds of *recording sessions* and what they imply for comparability to sections 3.2 and 3.3.

Several different notions of process comparability may pertain. Ultimately, it is up to the scientist to determine whether he wishes to include any two data objects in the same comparison in his study, and the computational histories in this model provide a good basis for that decision.

There are, however, several kinds of comparability that can be automated. For instance, two data objects may be comparable if their computational histories consists of exactly the same types of processes. For instance, if exactly the same smoothing or noise reduction algorithm was run on both datasets, it may not have reduced comparability. More generally we may not care about the exact algorithms used, but only their general class. Still more generally, we may care only about whether two computational histories share the same set of *inputs*. For instance, if one has been through a model-based analysis or been altered manually in a design process, it is unlikely that it is comparable to another which has not been affected by such an input.

3.1.5 Other Advantages

Note that the *experimental overview graph* provides a *vantage point* from which to consider the flow of data through an entire study, and which emphasizes the experimental design. Often scientists, when considering a study for metaänalysis, want such an overview— they wish to understand the methods employed, but they'd rather not look through the published paper to find such information.

The tracking of process and product information also furnishes a notion of *completeness*. In including data in an archive, we may well ask whether we have all the data necessary for replication of an experiment, or for metaänalysis. In this model we may simply ask whether the data form a fully-connected graph, with all inputs specified an all processes accounted for. If this is the case, then rerunning the analysis and reproducing a data object is only a matter of having the correct software.

3.2 Dimensional Vectors and Matrices

Individual *measurements* relate *quantities* with *scales* (see below). The problem of determining whether the data in two files are on a similar scale consists of: first, defining a representation of scale for measurements; second, generalizing that representation so that it applies to the theoretical (matrices, vectors, tensors), actual (recordings, images, volumes), and statistical (variance, expectation, principal components, Fisher's discriminate)

entities common in science; and third, using this generalization to decide a notion of comparability.

3.2.1 Nomenclature

The *scale* upon which a *measurement* is made specifies its *physical dimension* (i.e. mass, length, time) and uncertainty (including precision and accuracy). It also specifies other mathematical properties of the measurement *method*. For instance, whether the measurement device was restricted to *ordinal* measurement (i.e. greater than and less than), *interval* measurement with regard to a baseline (e.g. time and temperature), or the most general, *ratio* measurement. In methods which involve comparison to a baseline, that baseline must be characterized as part of declaring the scale.

A *waveform* is a series of measurements on the same scale, taken over time. A *tuple* is set of measurements in different scales, taken roughly at the same time. An *image* is a rectangular array of quantities, or *intensity values*, all on the same scale. A *volume* is the three dimensional version of the above. We also define *successive images*, which are videos, and *successive volumes*, such as fMRI scans.

These, collectively, are called *collections of measurements*.

We use *distributional effects* to mean the effects of mathematical transformations on both the uncertainty distribution of a measurement and the distribution of values in a collection of measurements.

Finally, an *interpretation* consists of a mapping from one scale to another. For instance, a recording device may receive information from its sensors in the form of analog variations in electrical current. Using an A/D converter, it can measure these variations and encode them with a certain degree of uncertainty. If the sensors are members of a CCD array, the device or its software applies an interpretation, mapping these measures of current onto measures of light intensity, and changing their physical dimension, uncertainty, and scale accordingly. Later, the scientist or his software may interpret these values yet again, perhaps as measures of dye concentrations at point locations on an organism, again transforming the scales as well as the quantities of the measurements involved.

3.2.2 Metrology and Mathematics

It is outside the scope of this paper to provide a formal framework for the application of notions of scale to linear algebra and statistics. We refer the reader to the excellent thesis by Tord Rikte, "On Physical Units in Mul-

tivariate Analysis",[13] G.W. Hart's book "Multidimensional Analysis",[14,15] and several papers by B. D. Hall, including "Software Support for Physical Quantities"[16] and "A Software Design Pattern for Handling Measurement Uncertainty".[17]

3.2.3 Capturing the Information

In section 3.1 we discussed the possibility of data management applications that retain information about processes as a scientist works. The same applications can associate information about scale, interpretation, and distributional effects with profiles of devices and algorithms. For instance, in addition to encoding the unix shell path and parameters of a processing technique, a software shell can also include information about the scale requirements for input files, any interpretations made, and any other transformations effected on units and scale by the algorithm.

Note that it is not generally necessary to actually implement statistical algorithms using software support for math with units, which could make them significantly slower, but merely to characterize the algorithms we use in terms of scale.

3.2.4 Comparability

The comparability of data depend on the statistical technique being used. For instance, while order statistics can be calculated for ordinal data, the mean and variance are meaningless. Similarly, eigenvalues are only meaningful in matrices for which the units follow a certain strict arrangement,[13] and the statistical techniques that rely on them (e.g. principal components) are thereby affected.

Two datasets with incompatible scales are not generally comparable before applying an interpretation that places them on the same scale. For instance, analysis via the general linear model[18] requires as input a set of factors that are assumed to (a) be on a linear distribution and (b) have a relation to the process being studied that is unmediated nonlinear functions. The act of assembling such an input usually requires the making of assumptions which can be expressed in metadata as interpretations.

3.2.5 Other Advantages

In a data model that supports the notion of interpretation as presented here with the explicit modeling of computational histories described in the previous section, we support an analysis of the graph structure of a study

which either accounts or does not account for all interpretations made. Many of these interpretations will not be questioned, will be standard across a field or research modality, and can be derived automatically by research software or by databases without troubling the user. Others really should be explicit and are an important part of evaluating the completeness, rigor, and statistical relevance of an analysis, both to the claims made by its author and to other claims.

3.3 Extant Phenomena at the Time of Recording

The problem of determining whether two sets of recordings (or derivatives of such recordings) are of comparable phenomena is inherently domain-specific, as the phenomena of study vary across domains. Domain *ontologies* for specifying this information are being created in every major field.[?,?,?,?,?] Here we consider a simple cross-domain structural mechanism for organizing such domain-specific information about the phenomena of study and for using it to decide comparability.

3.3.1 Nomenclature

A *recording session* unfolds within a certain *time span* and involves a set of *participants*, one of which is some kind of *recording device*. The "Run" object in Figure 1 is a recording session. Note that, although it unfolds in time, a recording session is considered a type of informational entity (e.g. a data object), rather than a process.

The *recording process* is the corresponding process: it takes the session as input and produces a data object containing just those aspects of the session that were recorded.

During the session, each participant undergoes a *participation process*, often directed by a *configuration*, a *protocol*, a *treatment*, or in the case of human participants, a set of *instructions*. Note that no *a priori* distinction is made between the participation process of the object of study and that of any other apparatus. This is because most recordings capture aspects of *both* the object of study and the apparatus, and are thus useful to several different audiences. For instance, in an fMRI scan participants may include a human subject following some instructions, a Sigma model scanner with a certain configuration, a display screen showing video, and a heart-rate monitor with its own configuration and setup.

3.3.2 Comparability

By avoiding the selection of a single object of study, and instead focusing on a set of participants, we both (a) widen the audience that might be interested in the study, and (b) offer an accounting for sources of noise and confounding effects in the recording session. The fMRI study in the example above, for instance, might be of interest to MR physicists and engineers interested in evaluating the effectiveness or information content of recorded material when produced by various models of MRI machines and various configurations. Similarly, a psychologist may wish to contrast the responses of fMRI subjects that were exposed to video with subjects who were not. In order to establish these as valid comparisons, it is important to provide a general framework within which the specifics of brain regions and tasks that might be present in a domain ontology can sit.

3.3.3 Capturing the Information

Satisfaction of this part of our metadata framework, unlike the others, may require annotations by the scientists themselves. Note that it should not be necessary for every experiment, as many experiments will likely have the same participants, although they may be following different scripts. Instead, manual annotation may be necessary during the process of experimental design innovation: it is new types of studies that will have different sets of participants in the recording session. This is perhaps acceptable, because it encourages the developers of new experimental designs and techniques to document possible confounding factors formally. In any case, if scientists wish to assess the comparability of older data with new recording techniques, they will need some kind of specification of what's changed..

3.3.4 Other Advantages

The interposition of the recording session into the production and analysis chain (see section 3.1) and its specification as a set of participation processes brings support to the visualization of studies and, as mentioned, the rigorous analysis of their data collection methods. In visualization, it is what allows us to represent an entire study as a single connected object. Note that the graph in figure 1 would be disconnected if it were not for explicit representation of the recording session.

4 Related Work

Due to its generality, this model has perhaps greater similarity to the scientific data formats (e.g. HDF,[?] NetCDF,[19] CDF,[?] etc) than to the various domain-specific metadata formats that have been developed. Although there are extensions to some of these formats for including data with physical units, none of them provide mechanisms for structuring the data in a study unambiguously. Indeed, these formats are made to be as loose and as general as possible, and thus they have many of the same limitations as the other loose formats of data exchange— text files, write-ups of various kinds, and semi-structured results databases (e.g. ACeDB,[20] PDB,[21] etc)— in that they accumulate mannerisms specific to each subfield and offer little overall structure recognizable to cross-disciplinary scientists or to general purpose software.

There are also (see section **??**) models that have been adopted by scientists from industry, including the very general metadata systems from library science and internet multimedia (e.g. Dublin Core,[22] MPEG-21,[23] MPEG-7.[?] While much good work has been done using business-driven models (these models do address many of the intellectual property, privacy, and access concerns in the scientific community), generally these models cannot accommodate the particular needs of science. The Indecs metadata framework,[24] which was originally developed to enable e-Commerce and rights management for multimedia data but has been adopted as the foundation for BioImage,[2,3] is one of the most powerful and extensive of these. It is a very well designed ontology, yet it is geared to annotation and distribution of end products rather than an open process.

Finally, we have domain-specific scientific metadata models, of which there is a greater variety than I was able to survey confidently. In every field there are several different and competing frameworks. In neuroscience we have NeuroNames, BrainML, NeuroCore, and the fMRIDC Ontology. Every large science database (e.g. GenBank, PDB, PubMed, etc) has it's own "next-generation" data format.

The BioImage ontology is one of the best that I surveyed: in addition to a detail domain ontology there are terms for expressing measured quantities with uncertainty and physical dimension, and there is some application of this to the notion of images and videos. However, these notions are not at present specified precisely enough to support a notion of completeness or statistical comparability.

5 Acknowledgements

The author would like to thank Jack van Horn for many intriguing discussions about information theory and statistical analysis, and for his emphasis on the open and complete sharing of data in the neurosciences. This work has also benefited substantially from various discussions and arguments with Jeff Woodward and Bennet Vance. Finally, without Mike Gazzaniga's unflagging support I would probably be bagging groceries.

References

[1] J.P. Birnholtz and M. Bietz. Data at work: Supporting sharing in science and engineering. In *Group 2003*, November 9–12 2003.

[2] Vivien Marx. IMAGING TECHNOLOGY: Beautiful Bioimages for the Eyes of Many Beholders. *Science*, 297(5578):39–40, 2002.

[3] Chris Catton. What is that crow thinking?: Separating fact and hypothesis in the bioimage database. In *E-BioSci/ORIEL Summer Meeting*, September 2–5 2003.

[4] Noor Jehan Kabani, David MacDonald, Colin J. Holmes, and Alan C. Evans. 3D anatomical atlas of the human brain. *NeuroImage*, 7(4):S717, May 1998.

[5] Joe Edelman. FCSM ontology. URL: http://orbis-tertius.net/x/fcsm, December 2003.

[6] fmridc home page. http://fmridc.org.

[7] Jason R. Swedlow, Ilya Goldberg, Erik Brauner, and Peter K. Sorger. Informatics and Quantitative Analysis in Biological Imaging. *Science*, 300(5616):100–102, 2003.

[8] James Hendler. COMMUNICATION: Enhanced: Science and the Semantic Web. *Science*, 299(5606):520–521, 2003.

[9] De Roure, N. Jennings, and N. Shadbolt. Research agenda for the semantic grid: A future e-science infrastructure. Technical Report UKeS-2002-02, National e-Science Centre, December 2001.

[10] David Huynh, David R. Karger, Dennis Quan, and Vineet Sinha. Haystack: a platform for creating, organizing and visualizing semistructured information. In W. Lewis Johnson, Elisabeth André, and John Domingue, editors, *Proceedings of the 2003 International Conference on Intelligent User Interfaces (IUI-03)*, pages 323–323, New York, January 12–15 2003. ACM Press.

[11] Joe Edelman. Hierarchical views for databases. Technical report, Dartmouth College, December 2003.

[12] M. A. Musen, R. W. Fergerson, W. E. Grosso, N. F. Noy, M. Crubezy, and J. H. Gennari. Component-based support for building knowledge-acquisition systems. In *Conference on Intelligent Information Processing*, 2000.

[13] Tord Rikte. *On physical units in multivariate analysis*. PhD thesis, Lund University, 2002.

[14] George W. Hart. The theory of dimensioned matrices.

[15] George W. Hart. *Multidimensional Analysis: Algebras and Systems for Science and Engineering*. Springer Verlag, 1995.

[16] B. D. Hall. Software support for physical quantities. In *Proceedings of the 9th Electronics New Zealand Conference*, 2002.

[17] B. D. Hall. Calculating measurement uncertainty using Automatic Differentiation. *Meas. Sci and Tech*, 13:421–427, 2002.

[18] Dale W. Stager. Sensitivity of the general linear model. Master's thesis, Washington Univ., 1969. Dept. of Applied Mathematics and Computer Science, 1969.

[19] Russ Rew and Glenn Davis. NetCDF: An interface for scientific data access. *IEEE Computer Graphics and Applications*, 10(4):76–82, July 1990.

[20] Richard Durbin and Jean Thierry Mieg. *ACeDB – A C. elegans Database: Syntactic definitions for the ACeDB data base manager*, 1992. `http://probe.nalusda.gov:8000/acedocs/syntax.html`.

[21] Helen Berman, Kim Henrick, and Haruki Nakamura. Announcing the worldwide protein data bank. *Nature Structural Biology*, 10(1):980, December 2003.

[22] S. Weibel, J. Kunze, C. Lagoze, and M. Wolf. RFC 2413: Dublin core metadata for resource discovery, September 1998. Status: INFORMATIONAL.

[23] Fernando Pereira. The MPEG-21 standard: Why an open multimedia framework? *Lecture Notes in Computer Science*, 2158:219–??, 2001.

[24] David Bearman, Eric Miller, Godfrey Rust, Jennifer Trant, and Stuart Weibel. A common model to support interoperable metadata: Progress report on reconciling metadata requirements from the dublin core and INDECS/DOI communities. Technical Report january99-bearman, D-Lib Magazine, January 18, 1999.