

Developing Metadata Standards for Scientific Data Reuse in NCSA's Distributed Grid Architecture

Joe Futrelle, National Center for Supercomputing Applications
152 Computing Applications Building, 605 E. Springfield, Champaign IL 61820
(217) 265-0296 futrelle@ncsa.uiuc.edu

INTRODUCTION

The unprecedented availability of network bandwidth and storage has brought about an explosion of on-line scientific data collections. In virtually all scientific fields, data is growing rapidly in volume and complexity. Managing distributed data collections is rapidly becoming one of science's key challenges, particularly for emerging interdisciplinary fields where it is critical for a given study to have access to multiple, heterogeneous collections[1]. Scientific data models are especially difficult to integrate because of their wide variety, high granularity, large storage requirements, and open-ended use requirements.

NCSA is fortunate to have an opportunity, through its Applications Technologies teams and its close relationships with government organizations such as NASA and the National Cancer Institute, to investigate new tools and strategies for integrating distributed scientific data collections. Through the development of application-specific use scenarios, we have attempted to sort out the common data modeling issues for a range of applications, including information retrieval, analysis, and visualization. In particular, we've been interested in metadata[2], since its primary use is to enable data access and interoperability.

ISSUES FOR SCIENTIFIC METADATA

We have found that the development of scientific metadata standards is as much a sociological as it is a technical challenge[3]. It requires new cooperative effort from communities of scientific data producers and consumers. These communities are much larger than teams working on a single collection, which is where most metadata development is currently taking place; however they are much smaller than entire scientific disciplines, whose data are typically much too diverse to be realistically integrated without glossing over significant detail. Building these communities is difficult, but in our view critical. Building consensus within these communities for metadata standards requires identifying target use scenarios and data collections, and agreeing on a consistent set of semantics across the target collections which meet the requirements of the use scenarios.

It is important to define scientific metadata standards in an implementation-neutral manner, with explicitly limited extensibility. Metadata standards that are tied to particular implementations are arduous to deploy, given many data providers' critical investment in specialized data management technology that supports important applications. Rather,

standards should be developed so that metadata can be encoded in various syntaxes and with various technologies and still retain its meaning; syntaxes are translatable only when the semantics of the source and target domain are known. Also, metadata standards that can be arbitrarily extended by individual data providers without community involvement can rapidly become useless as each data provider adds their own nonstandard features.

To the extent possible, metadata standards for one scientific community should be developed to interoperate at a general level with metadata standards coming out of other scientific communities. Ideally, a "common practice" can emerge out of the scientific community as a whole, enabling new kinds of interdisciplinary science.

DATA USE DRIVES METADATA DESIGN

Designing scientific metadata models is made difficult by the changing use requirements resulting from scientific and technological change. While it is impossible to accommodate all conceivable use scenarios, here are several key classes of scenarios that illustrate the issues involved.

Use Case: Retrieval

To use data, you must first be able to find it. Information Retrieval research has traditionally focused on text retrieval, but IR researchers and the digital library community are increasingly expanding their efforts to include other kinds of data, including scientific data.

The retrieval task consists of specifying criteria by which desirable datasets will be selected from a collection, and then browsing the selected datasets. For text-based retrieval, the selection criteria consist of keywords or phrases, which are matched against text indexes generated from a collection of documents. For scientific data, the selection criteria can be domain-specific, as can the indexing and matching procedures. Browsing results in a text-based retrieval system usually entails examining summary records consisting of information such as titles, keywords, and abstracts; for scientific data it can also involve examining low-resolution "thumbnails" of the scientific dataset, or statistics summarizing the data.

Retrieval across scientific domains requires metadata describing salient cross-domain features of the data. For instance, astronomy data might contain data about the abundance of particular chemical compounds in astronomical objects, whereas a biological database might contain

information about the metabolism of some of those compounds by organisms. In order to retrieve datasets relevant to a particular compound in both types of data collection, each needs to describe the compounds in the same way, or in a relatively unambiguously translatable way.

Furthermore, the data needs to be accessible within a shared context. For authors' names in a text database, this is rather trivial. But for scientific data, the issue becomes tremendously complex, since for example information about compounds might be encoded in the context of a spectrophotometer trace in an astronomy collection, but in a description of a metabolic pathway in a biological database. In many cases, context metadata critical for cross-domain retrieval is implicit in the collection that contains it. For instance each record in a botanical collection concerns plants, but few such collections explicitly encode that information with each record. This type of omission is logical for single collections, but limits the ability of data to be reused in a cross-domain, multi-collection retrieval scenario.

Use Case: Visualization

Scientific visualization increasingly means integrating a variety of data in order to show scientifically-salient relationships between them. To do this requires a unified spatial description of the various kinds of datasets to be visualized, and where the spatial characteristics of the data differ, meaningful ways to translate between the spatial representations. Metadata characterizing spatial representations can drive visualization applications capable of performing these translations, and producing integrated visualizations.

Earth science has played a leading role in developing standards for spatial representation. The OpenGIS consortium[4], NASA's EOSDIS[5], and research efforts such as VisAD[6] have all contributed to this emerging field.

One of the key challenges in the visualization of scientific data is the size and complexity of scientific datasets. Visualization applications are often computationally-intensive and may require special hardware, which means that they need efficient access to the datasets to be visualized. As scientific data increasingly becomes located in distributed, heterogeneous archives, it will become necessary to stage access to the data in order not to overwhelm visualization applications with unneeded data and network traffic.

NCSA has been working to address such problems by building an infrastructure for distributed management of computing and networking resources. The "Grid" will make it possible to perform visualization computations on remotely-located machines[7]. Data-driven visualization tasks such as subset selection will require that domain-specific metadata be transmitted from the user to the visualization application, so that it can perform

transformations of the visual representation of the data consistent with the scientific meaning of the data. For example, a visualization application showing some terrain needs to know that some data points represent snow cover in order to perform the visualization task: "show me the terrain without snow cover".

Use Case: Analysis

The term "analysis" covers an impossibly broad range of use cases for scientific data collections. However there are some general features of many analysis tasks that point to special metadata requirements:

- Non-interactive retrieval
- Integration across spatiotemporal scales
- Subsetting

Non-interactive retrieval means that analysis applications may need to access a retrieval system during the course of analysis in order to retrieve the data necessary to perform analysis steps. Traditional retrieval systems, which are geared towards interactive retrieval, are poorly suited for this kind of use. For example, web search engines, the technology of which has been widely adapted for scientific data collections, present non-standard HTML-based interfaces designed for humans to navigate but difficult or impossible to use from inside an analysis application. Non-interactive retrieval systems require well-defined retrieval protocols, query syntaxes, and metadata semantics.

Integration across spatiotemporal scale is as important for analysis as it is for visualization for many kinds of scientific data. For instance, NCSA's Chemical Engineering team has developed scenarios for the integration of codes based in continuum and non-continuum spatial scales in the problem of the electrochemical deposition of copper[8]. The NSF's Biocomplexity Initiative has stressed the need for integration of analysis across biological scales from the molecular to the ecosystem level[9]. In each case, the goal is to provide a framework whereby analysis on one scale can be correlated with the analysis on related scales, and that the appropriate computations can be performed to map data from one scale to another. To solve this problem in the general case, the spatiotemporal scale of data in a dataset must be explicitly described using metadata which a variety of scale-sensitive codes can interpret. A general-purpose spatiotemporal model such as the one implemented in VisAD[10] or the fiber bundle model developed by the U.S. DOE's ASCI program[11] can serve as the basis for this kind of metadata.

The ability to retrieve subsets of a dataset is critical for most analysis codes, which typically work on one part of a dataset at a time. Given heterogeneous underlying data representations, analysis codes must be able to express subsetting criteria in domain-specific semantics which are

translatable to the semantics of the variety of underlying representations. For example, an astronomical image processing code might extract regions from a sky survey in order to process them in parallel, even though the images in the sky survey might be represented in a variety of coordinate systems.

CASE STUDY: THE ASTRONOMY DIGITAL IMAGE LIBRARY

NCSA's Astronomy Digital Image Library[12] is a web-accessible repository of astronomical images, stored in the community-standard FITS format. Images are grouped into "projects", and metadata about the projects allow them to be retrieved based on a number of criteria ranging from authorship to astronomy-specific spatial coordinates.

The metadata associated with ADIL projects combines metadata derived from the FITS data and metadata submitted by the project investigators along with their images. The FITS-derived data includes spatiotemporal information. The investigator-supplied metadata includes the names of the observed astronomical objects, the names of the investigators, a project abstract, links to literature articles, and links to related projects.

Linking to literature is an important feature of the ADIL, since it allows researchers to locate the data associated with an astronomy article. This kind of linking is quickly becoming the new publishing model for scientific research, which will depend heavily on reuse not only of conclusions, but also of the data upon which the conclusions were based[13].

The ADIL's metadata is stored in a special-purpose Postgres database, but it is also exported in a general-purpose, reusable format called the Astronomical Markup Language[14]. AML is an XML document type designed to represent metadata about a variety of astronomy-related entities, including astronomical objects, people, articles, images, and data tables.

AML provides a portable, standards-based way of representing these objects for a variety of applications, including an AML browser and NCSA's Emerge[15] search toolkit. Standard XSL stylesheets[16] can easily be developed to reformat AML documents for a variety of applications, and perform tasks such as following links between objects.

Though the ADIL is a single collection, it has begun to address the need for interoperability within astronomy[17] by adopting the kinds of strategies and tools that are necessary to build interoperable services across multiple collections. The FITS format represents a consensus between a wide variety of astronomical practice; most domains have yet to reach such broad consensus on data representation. And AML is a

format which is carefully designed for reusability in a wide variety of contexts, including retrieval, browsing, and visualization. These are the basic building blocks upon which interoperable data services can be built.

- [1] R. R. Colwell. Testimony before the Senate Commerce, Science and Transportation Committee Subcommittee on Science, Space and Technology. 1 March 2000. <http://www.nsf.gov/od/lpa/congress/106/rc00301ngi.htm>
- [2] "NCSA Metadata Standards Working Group". <http://metadata.ncsa.uiuc.edu/>
- [3] R. E. McGrath. "Integrating Scientific Datasets and Digital Libraries", invited talk for the Center for Excellence in Space Data and Information Systems, NASA Goddard Space Flight Center, April 13, 1999. <http://www.ncsa.uiuc.edu/People/mcgrath/CESDIS/>
- [4] "Open GIS Consortium Home Page". <http://www.opengis.org/>
- [5] "NASA Earth Observing Systems Home Page". <http://eos.nasa.gov/>
- [6] W. Hibbard. "VisAD: Connecting People to Computations and People to People". *Computer Graphics* 32, No. 3, 1998, 10-12.
- [7] I. Foster and C. Kesselman, eds. *The Grid: Blueprint for a New Computing Infrastructure*. San Francisco: Morgan Kaufmann Publishers, 1999.
- [8] R. C. Alkire. "Electrochemical Deposition and Dissolution including Corrosion". http://metadata.ncsa.uiuc.edu/reports/alkire_991112.html
- [9] National Science Foundation. "Biocomplexity: Special Competition: Integrated Research to Understand and Model Complexity Among Biological, Physical, and Social Systems". Program Announcement NSF 00-22.
- [10] "Visualizing Scientific Computations: A System based on Lattice-Structured Data and Display Models". W. Hibbard, PhD Thesis, Univ. of Wisc. Comp. Sci. Dept. Tech. Report, #1226, 1995. <ftp://www.ssec.wisc.edu/pub/visad-1.1/lattice/>
- [11] "Accelerated Strategic Computing Initiative". <http://www.llnl.gov/asci/>
- [12] "NCSA Astronomy Digital Image Library". <http://adil.ncsa.uiuc.edu/>
- [13] R. E. McGrath, J. Futrelle, R. Plante, and D. Guillaume. "Digital Library Technology for Locating and Accessing Scientific Data", ACM Digital Libraries '99, August, 1999, pp. 188-194.
- [14] D. Guillaume. "AML (Astronomical Markup Language) Page". <http://monet.astro.uiuc.edu/~dguillau/these/>
- [15] "Emerge Home Page". <http://emerge.ncsa.uiuc.edu/>
- [16] "XSL Transformations". <http://www.w3.org/TR/xslt>
- [17] "The NASA Space Science Data System: A White Paper". July 1998. http://ssds.gsfc.nasa.gov/info/white_paper.html