SciFrame: a conceptual framework to describe data sharing in e-Science

Rodrigo Dias Arruda Senra, Claudia Bauzer Medeiros

[1]Institute of Computing – State University of Campinas (UNICAMP)
Caixa Postal 6176 – 13084-971 – Campinas – SP – Brazil

rsenra@acm.org, cmbm@ic.unicamp.br

Abstract. *The first SBC Challenge aims to provide solutions to the problem of managing large volumes of multimedia data. Our goal is to contribute towards research in these directions by discussing the problems involved in sharing scientific digital information. First, we propose a conceptual framework (SciFrame) that helps to understand the main issues involved and to integrate related research efforts. Second, we use a real case study to point out problems which are particular to scientific data management. Finally, we describe our case study using SciFrame.*

Resumo. *O primeiro Grande Desafio da SBC está voltado a soluções para o problema do gerenciamento de grandes volumes de dados multimídia. Nosso objetivo é contribuir para a pesquisa nessa direção através da discussão do compartilhamento de dados científicos em mídia digital. Primeiro, propomos um arcabouço conceitual (SciFrame) que ajuda o entendimento dos principais problemas envolvidos e integra esforços de pesquisa relacionados. Segundo, utilizamos um estudo de caso real para destacar problemas particulares ao contexto de gerenciamento de dados científicos. Finalmente, exploramos alguns problemas presentes no estudo de caso, organizando-os através do arcabouço proposto.*

1. Introduction

Computer Science has introduced a revolution in scientific research, and is recognized, nowadays, as being essential to the advance of science. The term *eScience* [Getov 2008] was introduced in the end of the 90's. While it originally focused in the computer simulations that accelerate scientific discovery, and the high performance distributed platforms these simulations ran on, it now encompasses several branches of Computer Science. Indeed, these platforms are fed by sophisticated instruments – e.g., telescopes, satellites, medical devices – which generate large volumes of complex and heterogeneous data at fast rates. These data should be processed by scientists using suites of complex algorithms and computational tools, and novel visualization methods. Interpreted results are fed back to the network, to become part of eScience data available.

As such, information representation and information sharing are both essential components of eScience. In fact, the World Wide Web – the most visible face of the Internet – was motivated by the need to communicate information among researchers. However, the ease to publish in the Web, associated with the vast amount of scientific data produced every day, caused an explosive growth in the amount of information available to scientists.

Voluminous data with a fast growth rate are just part of the problem. Heterogeneity is frequently cited as one of the most complex problems in data sharing. In eScience, it is aggravated by the inherent multidisciplinarity - besides the usual problems of variety in data acquisition, modeling, storage, processing and publication, all of which responsible for heterogeneity, there exists the issue that the scientists that participate in any given project have very distinct profiles and work contexts.

Another problem is how information is represented so that sharing can be facilitated. A white paper, a spreadsheet or a raster image are all valid representation formats, but not necessarily self-sufficient or complete. For instance, a white paper may lack details about the raw data used in an experiment, a spreadsheet may not inform from where or when the data were gathered, or a raster image might omit details about the sensors used for data capture. Completeness criteria depend not only on data producers, but also on the consumer's intent. This has prompted research on metadata, annotations and ontologies to enhance data characterization and provenance.

Sharing of data is just part of the problem – scientists also need to share *models*, which are defined in terms of sequences of operations, usually as scientific workflows – e.g., [Oliveira et al. 2008]. This paper does not directly cover workflows and models, concentrating on data aspects. We do, however, indicate several challenges associated with such workflows, which are closely connected with the second Grand Challenge of SBC – the management of models.

The goal of this paper is to exploit the many facets of the problem of sharing scientific digital data. This research is directly connected with the first Grand Challenge of SBC: management of large multimedia data volumes [Medeiros 2008]. Our main contributions to the Challenge concern the proposal of a conceptual model to be used as background to support an integrated analysis of these issues. Moreover, SciFrame can be used as high-level design pattern from which scientists can structure and describe their own work. The use of this model is exemplified through a real-world case of scientific data sharing. We conclude the paper with references to research efforts that try to tackle some of these issues.

2. SciFrame: A Conceptual Model to describe Information Sharing

According to Longworth [Longworth and Davies 1996], the stages in human learning can be described by the following *information ladder* also known as the *DIKW model*: $Data \rightarrow Information \rightarrow Knowledge \rightarrow Understanding \rightarrow Insight \rightarrow Wisdom.$

While the rightmost stages belong to the domains of cognition, psychology and philosophy, the first three steps are directly related to the first SBC Grand Challenge, and to SciFrame. The terms *data*, *information* and *knowledge* have various definitions and can be used for overlapping concepts. However, in the context of SciFrame, we adopt the following definitions: **Data** is a structured collection of *typed values*, represented in *digital form*. The important distinction between *data* and *information* is that the latter has **explicit** semantics. **Information** is a set of inter-related data, *bound to semantics* and *useful for some purpose*. From a Semiotics point-of-view, data are symbols and

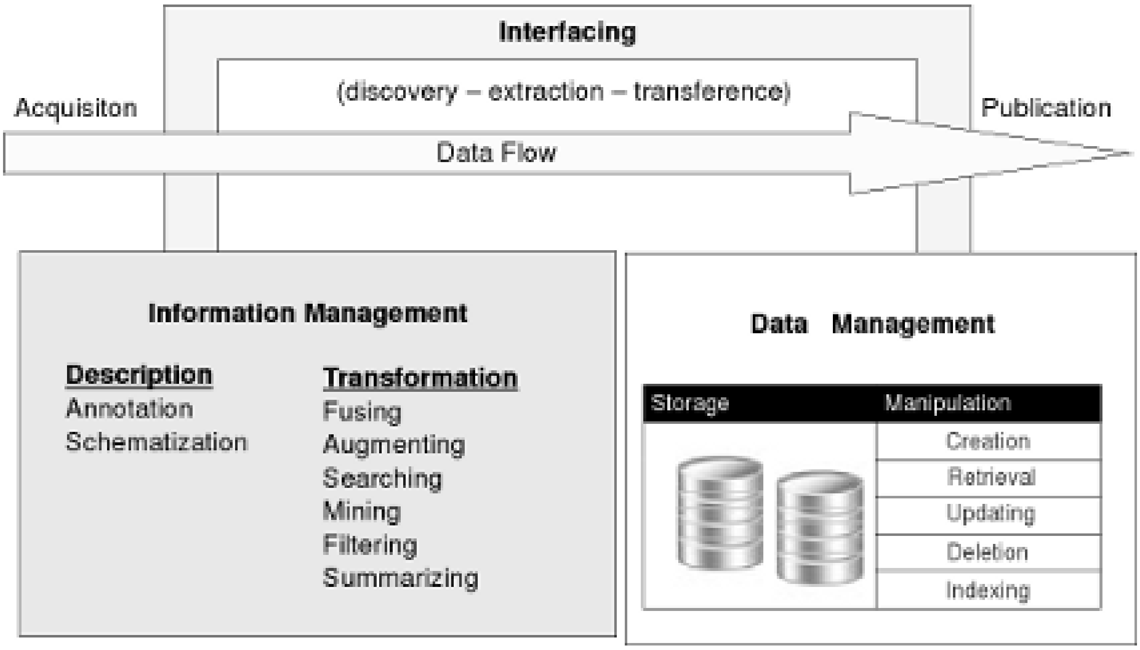

Figure 1. SciFrame - Scientific Digital Data Processing Framework

information occurs when data are used to refer to something. **Knowledge** represents the *cognitive dimension* of the information's consumer, linking information from the process domain to information present in the "out-of-process world". From the DIKW model, knowledge is created by *using the information for action*.

In this section we define a conceptual framework that describes systems or processes involving scientific digital data manipulation – *the Scientific Digital Data Processing Framework*, or **SciFrame** for brevity. SciFrame is depicted in Figure 1, and it is divided into three high-level abstractions: Interfacing, Data Management and Information Management. This overall structure and its elements are well known and compose a long-time adopted pattern. Nevertheless, this pattern lacked a cohesive definition in a standardized vocabulary, that we try to remedy here.

Let us consider an eScience process that involves some kind of scientific data manipulation. **Interfacing** defines the process boundaries and comprehends all digital data exchange between the process itself and external entities, either human or artificial. The *Interfacing* element has two subdivisions: *acquisition* and *publication*. **Acquisition** represents the obtention of data: in digital form, with a known structure, from a known source, through a given media. This stage represents the process input. **Publication** stands for suitable data representations that allow proper communication with external entities. Data are published in digital form, with a known structure, via a given media. This stage represents the process output. The *acquisition* and *publication* elements of a given eScience process are potentially independent from each other. When two processes are interacting, the acquisition element of one is coupled to the publication element of the other [Pastorello Jr et al. 2008].

One way to define their data exchange pattern is to determine which role takes initiative in the transaction. A *PUSH* pattern occurs when the data provider initiates the data exchange, and conversely a *PULL* pattern occurs when the data consumer takes the initiative. Data exchange can be divided into three stages: discovery, extraction and transference. **Discovery** represents the acquisitor's concern with identifying suitable data publishers given a set of information needs. On the other hand, publishers are concerned with making themselves well-known to potential acquisitors. **Extraction** represents the problem of extracting the information from a chosen data provider/publisher. **Transference** represents the problem of moving data from the publisher into the data management facilities of the consumer.

SciFrame makes the distinction between *Information Management* and *Data Management*. *Information Management* is responsible for higher-level information manipulation, analysis and synthesis. *Data Management* is responsible for lower-level data manipulation for persistence purposes.

The **Data Management** element is subdivided into *storage* and *manipulation* elements where: **Storage** is responsible for data persistence (caching inclusive). **Manipulation** provides support for the basic interactions with the storage element. These interactions are called CRUD, an acronym for the operations: create, retrieve, update and delete. We have also included the index interaction. Therefore, we shall refer to it as CRUDI.

In the context of *Information Management*, an eScience process can be examined and modeled according to two main axes: *data description* and *data transformation*. **Transformation** corresponds to a finite number of operations that the process applies to the stored data, in order to change its contents or structure. We present an informal set of definitions for the operations. **Augmenting** adds information to the data present. **Fusing** generates new information by coalescing part of the data present. **Filtering** decreases the amount of information by discarding data. **Summarizing** decreases the amount of data by classifying, clustering or generalizing data. **Searching** locates information inside data. **Mining** extracts unperceived information from data.

The last two, searching and mining, can be seem as idempotent transformations, considering that the status of the data is not altered. Transformation can occur at any time, but when it takes place immediately after acquisition, it is often called pre-processing. The *Description* element is orthogonal to the *Transformation* element. It corresponds to the gathering and organization of information about the process data elements, documenting their nature, structure and purpose. It is also encompasses the roles of *annotation* and *schematization*. As an example, *provenance* is one of the most important types of description, fundamental in the eScience context to ensure the shared data elements' quality and usefulness.

We acknowledge that SciFrame requires a more complete and formal characterization of the interactions and dependencies of its constituent elements. However, due to restrictions in this paper's length, we chose to present a case study instead. The case study illustrates a typical example of a scientific application with a strong focus on information and data sharing, evidentiating the role of SciFrame as a generic pattern to describe eScience research.

3. Case Study: Crop Monitoring in WebMAPS

In order to illustrate SciFrame's applicability, we present a real-world scenario from the WebMAPS eScience project [Macario et al. 2007]. This is a multidisciplinary effort in-

Data Management		
Storage		organize and persist input raster images and their textual metadata
		organize and persist composite NDVI profiles
Manipulation		index and fetch regions from images
Information Management		
Description		spatial regions of interest are described in Well-Known Text notation (WKT), raster images in HDF format have embedded textual metadata
Transformation		eliminate clouds by generating composite images (data gaps removal)
		detect and mitigate perturbing factors
		calculate NDVI time series
		filter out noise using HANTS (Harmonic Analysis of Time Series)
Interfacing		
Acquisition	Discovery	elect adequate remote sensing data products and providers available in the Web
	Extraction	identify a path to data products in the provider's Web portal
	Transference	download products (raw multispectral satellite images) via HTTP or FTP
Publication		publish NDVI profiles as 2D scatter plots (average NDVI vs time) in WebMAPS portal

Table 1. NDVI profile generation described with SciFrame

volving computer scientists and experts on agricultural and environmental sciences to develop a platform for agro-environmental planning. The case study concerns an eScience process within WebMAPS, and shows that sharing scientific data presents challenges that are not found in other kinds of data sharing (e.g., in industrial or business contexts).

An important problem in agro-environmental planning is monitoring crop behavior. One of the earliest studies [Ulaby 1975] about deriving crop condition from solar radiation has shown that there is a strong correlation between radar measurements (backscatter) and *leaf area index* (LAI). LAI determines the amount of energy available to the plant for photosynthesis which in turn drives the plant development and subsequent yield.

One of the tools used by experts to monitor crop behavior is based on *Normalized Difference Vegetation Index* (NDVI). Informally, this index represents the healthiness ("green-ness") of a given vegetation cover. It is computed as the difference between the red (RED) and near-infrared (NIR) bands of multispectral images, given by the formula: $NDVI = (NIR - RED)/(NIR + RED)$ There are several vegetation indexes proposed, such as: Perpendicular Vegetation Index [Richardson and Wiegand 1977], the Soil-Adjusted Vegetation Index (SAVI) [Huete 1988], the Atmospherically Resistant Vegetation Index (ARVI) [Kaufman and Tanre 1992] and the Global Environment Monitoring Index [Pinty and Verstraete 1992]. Choosing amongst them is part of the problem – distinct scientists favor different indexes, which result in incompatible analyses. However, NDVI remains the most well-known and used index to detect live green plant canopies from multispectral remote sensing data.

One of the processes in WebMAPS corresponds to a tool that generates *NDVI profiles*. Each profile is a time series of average NDVI measurements, which represents the vegetation's health (biomass status) in a particular region for a given time period (crop's phenological cycle). NDVI profiles characterize the spatio-temporal behavior of specific crops. They allow experts to monitor the evolution of the crop, detect (and prevent) anomalies and forecast crop yield.

The management of spatio-temporal data series is a problem common to many eScience domains, which is one of the reasons for our choosing this case study. Table 1 gives an overview of the process that generates *NDVI profiles*, summarized under our

SciFrame conceptual model. Some processing details were omitted, but it serves the purpose of illustrating SciFrame's application.

For instance, although the NDVI formula is mathematically simple, satellite image pre-processing is complicated and requires extensive data correlation. Perturbing factors should be detected and mitigated in order to avoid negative influence in the computed NDVI. They include: (i) high level of water vapor and aerosols; (ii) soil moisture; (iii) angular geometry of illumination and observation at the time of the measurements; (iv) sensor-dependent data calibration. These issues represent nested processes in NDVI profile generation. Though not described in this paper, we point out that they have a SciFrame's description of their own.

3.1. Practical Pitfalls

Consider a scientist in charge of studying sugarcane crops in Ariranha County in São Paulo state (Brazil). The goal is to analyze sugarcane yields using year 2001 as benchmark, when it was the top producer county, yielding approximately 5.15 million tons/year. This scientist decided to use a NDVI profile as an estimator [Carlson and Ripley 1997], based on previous studies of NDVI correlation to crop yields.

In order to do that, first of all, this scientist needs the georeferenced perimeter of every farm growing sugarcane in Ariranha County. Georeferencing means to establish an appropriate set of coordinates defining accurately the region's location on the Earth's surface. Here we face a common barrier in eScience – data availability. Georeferenced boundaries may be hard to obtain in practice, due to the lack of reliable boundary databases. This may be circumvented by a ground survey with GPS measurements, which requires the farmers agreement. We are not concerned here with confidentiality issues. Data privacy and security are valid open problems in eScience data sharing that are out of the focus of this proposal.

In addition to the spatial regions of interest, the scientist must collect remote sensed imagery covering the county's area (aprox. 133 km^2) during the target years. NASA's MODIS sensor is a reasonable data source. One of its derived products is "MOD13Q1 - Vegetation Indices 16-Day L3 Global 250m". This dataset is delivered by NASA already pre-processed, with 11 pre-calculated vegetation indexes, including NDVI. MOD13Q1 is distributed as files encoded in NASA's HDF-EOS format [Qu et al. 2007], each covering an area of 5,760,000 km^2 with average size of 500 Mbytes.

This means that a 2-year long NDVI time series covering Ariranha County's area (532 pixels per MOD13Q1 image) requires 46 images. The data volume amounts to roughly 47.8 Kbytes. However, due to distribution granularity (in 500 Mb files), this scientist will have to download 23 Gbytes. Therefore, 99.99979% of the downloaded data is useless considering the target study. In the worst case where a single satellite image snapshot (swath) does not entirely contain the region of interest, the waste is doubled.

Moreover, to download the 46 files (92 in the worst case scenario), the scientist will have to fill out a NASA web form (maybe several times) specifying product, swath (region of interest) and time interval. The estimated delivery delay can range from a day to a week, depending on the available network throughput. Nevertheless, that is still a straightforward process. Each web portal providing remote sensed imagery implements a different acquisition workflow. Some portals demand a round of e-mail exchanges prior to data release. Other portals arrange the files in hierarchies, forcing the user to browse through several pages before reaching the target links.

After all data are obtained, there remains the issue of compatibility with the scientist's processing environment. For instance, computing time and storage space may be required to convert NASA's HDF-EOS format into more widespread input formats such as GeoTIFF [Ritter and Ruth 1995], NetCDF [Rew and Davis 1990], HDF4, or HDF5 [Folk et al. 2002].

Once the satellite data are converted and stored, several other issues remain. For instance, the acquired images may present gaps in the region of interest (e.g., clouds) that could be compensated by additional processing (e.g., by acquiring images from other providers, or executing complex data manipulation procedures). Image noise has to be taken into account.

Once all preprocessing is finished to the scientist's satisfaction, the profiles can be generated. Again, this presents many challenges. For instance, just as they may choose distinct vegetation indexes, research groups may adopt different procedures to generate profiles, which in turn may result in differences in profiles. Notice that each processing strategy chosen will compound the obstacles to sharing profile data. Hence, in order to share the published profiles with other groups, an appropriate description of the entire profile generation process must be provided – e.g., indicating the source images used, the scientific workflow selected to create the profile, and so on. This kind of discussion falls into the general problem of *provenance* in eScience.

These are just a small sample of problems involved in sharing eScience data.

4. Conclusions

The first Grand Challenge of SBC involves the management of multimedia data, which includes scientific data. There are many issues concerning the latter that need to be investigated using specific procedures, given some of their peculiarities. This paper is a step towards this direction, providing an integrated perspective of efforts and phases involved in sharing of eScience data. The SciFrame model, conceived with this in mind, was presented through a case study of scientific data manipulation in WebMAPS. This study illustrates some typical problems related to data sharing, particularly the problem of distributing large datasets over the Web. We concluded the paper pointing out that SciFrame should be used as a design pattern, from which scientists could structure and describe their own eScience efforts.

Acknowledgements

The authors thank FAPESP-Microsoft Research Virtual Institute (eFarms project), CNPq and CAPES for the financial support for this work.

References

Carlson, T. and Ripley, D. (1997). On the relation between NDVI, fractional vegetation cover, and leaf area index. *Remote Sensing of Environment*, 62(3):241–252.

Folk, M., McGrath, R., and Yang, K. (2002). Mapping HDF4 Objects to HDF5 Objects. *National Center for Supercomputing Applications, University of Illinois, July*.

Getov, V. (2008). e-science:the added value for modern discovery. *Computer*, 41(11):30–31.

Huete, A. R. (1988). A soil-adjusted vegetation index (savi). *Remote Sensing of Environment*, 25:295–309.

Kaufman, Y. J. and Tanre, D. (1992). Atmospherically resistant vegetation index (arvi) for eos-modis. In *Proc. IEEE Int. Geosci. and Remote Sensing Symp*, pages 261–270.

Longworth, N. and Davies, W. K. (1996). *Lifelong Learning*. Kogan Page Ltd.

Macario, C. G. N., Medeiros, C. B., and Senra, R. D. A. (2007). O projeto webmaps: desafios e resultados. In *IX Brazilian Symposium on GeoInformatics - Geoinfo 2007*, pages 239–250, Campos do JordÃ£o - SP. INPE.

Medeiros, C. B. (2008). Grand research challenges in computer science in brazil. *IEEE Computer*, 41(6):59–65.

Oliveira, F. T., Murta, L., Werner, C., and Mattoso, M. (2008). Using provenance to improve workflow design. In *Provenance and Annotation of Data and Processes: Second International Provenance and Annotation Workshop, IPAW 2008. Revised Selected Papers. LNCS*, pages 136–143.

Pastorello Jr, G. Z., Senra, R. D. A., and Medeiros, C. B. (2008). Bridging the gap between geospatial resource providers and model developers. In *GIS '08: Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, pages 1–4. ACM.

Pinty, B. and Verstraete, M. M. (1992). GEMI: a non-linear index to monitor global vegetation from satellites. *Plant Ecology*, 101(1):15–20.

Qu, J. J., Gao, W., Kafatos, M., Murphy, R. E., and Salomonson, V. V. (2007). *Earth Science Satellite Remote Sensing*, volume 2, chapter The NASA HDF-EOS Web GIS Software Suite, pages 245–253.

Rew, R. and Davis, G. (1990). NetCDF: an interface for scientific data access. *Computer Graphics and Applications, IEEE*, 10(4):76–82.

Richardson, A. J. and Wiegand, C. L. (1977). Distinguishing vegetation from soil background information(by gray mapping of Landsat MSS data). *Photogrammetric Engineering and Remote Sensing*, 43:1541–1552.

Ritter, N. and Ruth, M. (1995). GeoTIFF Format Specification GeoTIFF Revision 1.0.

Ulaby, F. (1975). Radar response to vegetation. *IEEE Transactions on Antennas and Propagation*, AP-23(1):36–45.