SCIENTIFIC DATA MANAGEMENT: A SURVEY AND RESEARCH DIRECTIONS

Rehman M. Abdul

Department of Computer Science, Sukkur IBA-Institute of Business Administration,
Sukkur, Pakistan rehman@iba-suk.edu.pk

Summary

In today's electronic world managing scientific data, especially in natural sciences domain, has become more and more challenging for domain scientists since the data stemming from scientific experiments are extremely heterogeneous in terms of structure (schema) and semantics (interpretation). Furthermore, the involvement of increasingly large number of scientific instruments such as sensors and machines makes the scientific data management even more challenging since the data generated from such type of instruments are highly complex. In order to address the data management issues, in past years, many data management / integration systems in diverse scientific domains have been emerged. However, these systems are not providing ideal solutions for many of the data related issues. Moreover, due to the high level of diversity in specifications, standards and frameworks from diverse application domains, selecting a suitable system for a particular data management task is not always an easy process. As a solution to this common problem, this paper presents a concise survey of existing data integration systems from the scientific domain and makes a number of key suggestions towards the future development of data management systems.

Keywords: Scientific data management. Scientific applications. Data integration. Data transformation. Tools.

1. Introduction

Data management in scientific applications is not a "one-shot approach"; rather it encompasses multifaceted tasks of data management at different levels. At the syntactic level, assorted data formats need to be handled since scientific data come in diverse representational formats such as PWD, HDF, HDF-EOS, netCDF, TIFF, GeoTIFF and so on. At the logistic level, the physical movement of the data between heterogeneous data sources is of significance since data sources are geographically distributed, under different platforms, employing different communication protocols and offering different access methods. At the structural and semantic levels, data stemming from heterogeneous sources need to be integrated and transformed structurally as well as semantically. Also, data integration in scientific domains such as environmental sciences is no more the matter of just combining the data horizontally and vertically, resolving semantic conflicts on an equivalence basis, and filtering out the values, but go beyond that since sensor and machine generated data are mostly raw, unstructured or even missing; these need to be validated, interpolated, and aggregated. Data management / integration systems, in scientific applications, have emerged as a principal approach to tackle such sort of issues. So far, several approaches have been developed and published (see [1, 12] for different surveys). In order to define data management / integration task, most of these systems offer either query-oriented or portal based interfaces. Some of them also provide the means to define the data management task in terms of data operations in a pipelined (workflow-based) fashion. However, our real world experience in diverse scientific domains [13, 9] convincingly demonstrates that despite the availability of a wide range of systems, the domain scientists are feeling uncomfortable in managing their data. This is due to the two main reasons. One: insufficient experience and familiarity of the domain scientists towards data management technologies that already exists. Two: the diversity in specifications, standards and frameworks from multiple domains, which results in ambiguity in taking decisions about the choice of right / suitable tool for a particular task. In this paper, we present a concise survey of prominent data management / integration systems from different scientific domains. Moreover, based on the survey we also provide some key suggestions towards the future development of the systems.

2. Data Integration Systems

In this section we discuss some prominent data management / integration systems in different scientific disciplines. Conventionally, the main objective of data integration system is to reconcile a number of heterogeneous data sources by providing a uniform access interface. In this way, these systems attempt to keep users (domain scientists) away from the implementation details of how the data are structured at individual sources and how they are to be integrated. In order to reconcile diverse data sources two approaches are usually followed; warehousing and mediator-based integration [5].

LifeDB [6**]** is a wrapper (mediator) oriented data integration system that falls under the category of semantic mediation and uses ontologies for semantic transformation, supporting concept, attribute and value equivalence. Main components of the system are the schema matching subsystem called OntoMatch, the wrapper generation subsystem known as FastWrap and the Integration subsystem. OntoMatch subsystem provides schema mediation, by taking two relational schemas as input and returns schema correspondences as a list of pairs of equivalent attributes. The matching criteria is exclusively based on equivalence, thus it does not consider concept subsumption, generalization, specialization, and navigation. In order to extract data

from multiple sources (local or remote), the system generates source specific wrapper which provide the access interface to the source. Once data are extracted from sources and transformed into common data model backed by RDBMS (MonetDB [22]), the sub-system performs integration. The key advantage of the system is the utilization of two separate sub-systems for two different operations, making the system flexible since it promotes the principle of modularity. The system also offers a declarative query language known as Bioflow [14] for specifying integration task. The specification of the data products on physical level (user needs to specify data objects concretely with the schema specification) and definition of data operations through low level SQL-like fashion make the system harder for domain scientists to follow since scientists are not database experts, thus require proper abstraction.

Pegasys [18**]** is a mediator-based software system that aims at executing biological sequence tools, defined in pipelined (workflow-based) fashion. Not only this but also it offers the integration of analysis results stemming from different tools. The workflows can be created using a graphical user. It offers several services for job scheduling, execution, database interaction, and adaptors. The main functionality of application layer is to convert the workflow (defined in XML) into a directed acyclic graph (DAG) of analyses tasks. The results stemming from analysis are inserted into the backend database layer with help of specific adaptor depending on the type of analysis tool, thus the data exchange between two analysis tools is done by invoking at least two adaptors; one for selecting (also integrating) the data from output of one analysis and importing into common data model whereas another adaptor for exporting the data from backend database layer. Data conversion, integration and transformation (interoperability) are done under the hood of an adaptor that is specific to the application. Integration specifications (in terms of queries over the source datasets) are hard coded into adapter implementation, nevertheless system supplies a bunch of pre-configured adaptors for exporting data in GFF and GAME XML and for importing into Apollo genome editor [16] format.

In **HyperGrid [**21**]** a scientific data processing framework, especially for environmental sciences, known as *HyperGrid* is presented which falls under the category of mediation application, but not in the semantic mediation, i.e. system does not offer any kind of semantic interoperability. The system comprises new data model (inspired by grid data structure), query processing framework and several generic optimization techniques. The main motivation of the system is to prepare (referred as "Data preparation" step) datasets stemming from sensors before being used in any analysis. Once raw sensor readings are wrapped up into the common data model and then specified operations are performed. The key points of the system include the support for several advanced query optimization techniques.

SEEK [15, 8**]** (Science Environment for Ecological Knowledge) is a framework which is especially designed to facilitate scientists from ecological domain in managing and integrating their data. The architecture of SEEK is divided into three-layers: *EcoGrid, Kepler* (a Scientific Workflow System), and *Semantic Mediation System* (SMS). The separation of these layers is beneficial in the sense that application related issues are handled at topmost layer via the *Kepler* system and the data management issues are dealt at lower layers via *SMS* and *EcoGrid* systems, reducing scientific workflow complexities. In SEEK, data management specification is encoded into the implemented application under the hood of SMS middleware and a component view is provided to the upper layer (Kepler workbench). Then the end-users employ that component as a processing step at the application workflow layer. The ideal approach is to follow top up approach since in this approach the data management specification is defined at abstract level at upper layer (in a separate workflow) and then sent to the lower layer for the execution. The main benefit of this approach is that the specification is not encoded under the application stack, rather configurable at abstract level.

TAMBIS [11**]** (Transparent Access to Multiple Bioinformatics Information Sources) is a mediator-based software system. In order to define the queries TAMBIS provides a graphical user interface in which users are required to browse through concepts and select the ones that are of interest. These concepts are defined in terms a global schema. The queries are first specified through a graphical query language so called GRAIL (declarative source-independent description logic). This GRAIL query is then transformed into an internal representation so called Query Internal Form (QIF). Finally, QIF is converted into a query execution plan (in terms of wrapper) in CPL (Collection Programming Language) program. Queries expressed in GRAIL just specify "what" is required. They do not provide information about "how" and "from where" the request has to be served. This information is provided by the query planning and transformation layer which is encoded already. In order to access the underlying sources TAMBIS uses wrappers from the BioKleisli [10] system. No doubt that the system facilitates domain scientists since in order to analyze and integrate the datasets they always traverse through the multiple biological and bioinformatics. However, due to the main focus on providing a global view over multiple sources, the system is not utilizable in ad hoc based data exchange scenario where data sources have not always the same view.

BACIIS [4**]** (Biological and chemical information integration system) is a mediator-based information integration software system, with main focus on the reconciliation of life science web- databases. BACIIS follows the mediator-wrapper approach, thus each data source participating in the integration is associated with specific wrapper. Instead of creating a global data schema, in BASCIIS each source schema is mapped onto

domain (global) ontology. In order to manage different formats, it is the responsibility of mediator-wrapper to convert the representational format of the data in the source database to the internal format used by the system. The system is almost analogous to TAMBIS, thus the main focus of the system is on providing a global view over multiple heterogeneous data sources and allowing the users to traverse through global schema. Like TAMBIS, it provides a web-based common interface that offers to browse through schema terms and their properties and also to add some filtering conditions. The approach gives the impression of being a portal. The integration is limited to combine the sources vertically and horizontally. The system follows a semantic mediation for data integration and extracts the data via the implementation of specific wrappers. However, the system is to provide global view over a number of local views of different data sources, thus the system is not utilizable in ad hoc based data exchange scenario where data sources have not always the same view.

BUSTER [20**]** (The Bremen University Semantic Translator for Enhanced Retrieval) is mediator-wrapper oriented system and follows semantic mediation. Its goal is similar to BACIIS that is integrating diverse data sources and producing results as integrated views. Unlike BACIIS and TAMBIS, its architecture introduces two phases, namely Acquisition and Query. The acquisition phase aims at acquiring all the necessary information needed for a data integration task. This mainly includes source specific information so called Comprehensive Source Description (CSD), which is used for identifying the data sources, and the specification of data integration task so called Integration Knowledge (IK), which describes how the information can be transformed from one source to another. The architectural components of the system are classified into three levels: the

syntactic, the structural and the semantic. On syntactic level, wrappers are used whereas on structural level, specialized mediator components, which are configured by transformation rules, are utilized. On the semantic level, two specialized tools are exploited, i.e. functional context transformation (for simple mapping based) and context transformation by reclassification (for complex transformation involving generalization and specialization). As per data management issues at multiple levels, the system manages different issues of scientific data management on different levels. However, the system does not focus on operational level that deals with the issues of domain specific data processing and data logistic level that is responsible for data transportation (physical movement of the data) issues.

ALADIN [3**]** (ALmost Automatic Data INtegration) is wrapper oriented data integration approach in life sciences using warehousing technique, thus it falls into the category of "warehousing based integration". As a basic data model, ALADIN uses a relational database, thus it is limited to the management / integration of datasets that can be converted to a relational representation, including XML data and flat-file databases. In order to perform integration the system uses the techniques from schema matching [17] and data and text mining [7]. Following the warehousing integration technique has some benefits in scientific applications, for instance data are available locally thus many network as well as source related problems can be eliminated, for instance network bottlenecks, low response times, and the unavailability of sources. However, warehousing technique (Clean-Store-Query) is not feasibly in situation where data sources are frequently changing and also where always current information from data source is expected such as ad hoc data exchange, requiring data operations to be performed during the query execution.

SnapLogic [19**]** and **Apatar [**2**]** are wrapper-oriented data integration and ETL (Extract, Transform and Load) tools at enterprise scale. The systems offer a nice graphical user interface that allows users to graphically design and execute the data integration tasks. In order to design integration pipelines (workflows) the systems are supplied with a rich number of built-in and pre-configured components. Data sources are treated as instances of these components and can thus be customized graphically. The advantageous aspects of these systems include very nice user-friendly graphical interface with very rich libraries of data operations. However, in the integration pipeline, sources and operations determine the physical design, thus no conceptual separation between the logical specification of the operations and their physical implementation can be achieved. Furthermore target community of these systems is database community, thus data model is overwhelmed by record orientation, making it challenging to incorporate other scientific data formats.

3. Discussion

We presented an overview of some prominent scientific data integration systems. Through our experience of evaluating these systems, we make some key observations for improvements and future research directions, by considering three criteria, i.e. usability, sustainability and adoption of the tool.

Keeping in mind the expertise and knowledge of domain scientists towards data management technicalities, the ideal system should separate the specification of data processing from its physical implementation, empowering scientists to specify "**what to do**" at abstract level while the system resolves the details of "**how to do**" at physical level. This gives high level of freedom to scientists, allowing them to focus on domain challenges in the realm of scientific experiments. It has been observed that in most of the above cited systems data management specification is defined at physical level, ignoring or paying less attention

towards the design of such specification at abstract level.

Our experience in diverse scientific domains demonstrates that extensions to the existing system are often required as new scientific machineries as well as datasets are rapidly introduced. Modularity in the architecture will help in sustainability and adoption of the system in new and changing scenarios, as it will transparently directs **how** (the way) to add **what** (new functionality) and **where** (in which module). Scientific experiments are assumed to be the series of analytical steps which often consume and produce huge amount of heterogeneous and distributed data objects relevant to their current study. These data objects can be primitive or complex type, files in different formats and sizes, database tables, or other forms. Without proper abstraction, scientists are often overwhelmed and lost in the sea of heterogeneous and distributed data objects. Hence, at the application level an abstraction which hides the technicalities of physical data objects is of extremely importance for domain scientists. Most of the previously mentioned systems work on the level of files. In some systems, the data objects are defined through the specification of a concrete entity (that asserts a physical database table) with its concrete structural specification. Some systems also enforce the data objects to be in its proprietary data model that is based on a grid data structure. Some offer the functionality to access the data products via a single global view. Accessing the data via a common interface helps a lot to abstract many data management operations, but data products and their preparation logic are encoded into the software stack thus cannot be defined and customized on abstract level.

In the most of these systems data formats are managed through the implemented wrappers which convert the format of the extracted dataset to the internal representational format. Another approach is to manage the formats out of the wrapper implementation and done after the data extraction through an explicit conversion step. We believe that advantages from both the approaches can be achieved through combining both the approaches. This means that formats have to be dealt during the extraction operation but the operation should be customizable and configurable (i.e. logic should not hard coded into wrapper implementation) at abstract level so that the assorted formats can easily be incorporated systematically.

Structural and semantic compatibility lays the foundation for the data integration. No doubt that structural compatibility has much been focused by these systems whereas semantic compatibility has also been worked out by some. The main problem in this area which we experienced when we were working with domain scientists is that they are unable to define semantic descriptions such as ontologies, mappings and correspondences. Just to tell them they need to use the ontologies is not enough helpful for the scientific community. They need a proper guidance and integrated tools that facilitate them to structurally define and maintain the ontologies, schemas, mappings and so on. No doubt that the above mentioned systems also focused on semantic perspective of scientific data, however the most challenging aspect of semantic interoperability especially in scientific applications is the proper guidance for management of semantic descriptions. Thus, the ideal framework should also provide the clear guidelines such as: who will define the ontologies (formal semantic description) and how? How to maintain the evolution of ontologies?

Data logistic between distributed systems is the common characteristic of scientific applications and it lays the basis for the real data exchange. Due to the central focus on the data integration by most of the above mentioned systems, physical data transportation is not much thought- out. Fundamentally, data logistic can be achieved by employing two operations, i.e. extracting data from one system and loading into another one. In fact this issue has been mainly focused by **ETL** community where **E** (Extraction) and **L** (Loading) can be assumed to be the logical solution for the physical data movement between the data sources, while traditional integration systems exclusively work on **T** (Transformation).

4. References:

1. Agrawal R., Rege M. Data integration from scientific data sources: A Survey. in Proc. of 7th Int. Workshop on Computer Science and Information Technology CSIT'05, Uba, Russia, 2005

2. Apatar: An open source data integration and ETL tool (home page), www.apatar.com, [5. 2011]

3. Bauckmann J. Automatically Integrating Life Science Data Sources. In VLDB PhD Workshop, 2007

4. Ben-Miled Z., Li N., Baumgartner M., Liu Y. A decentralized approach to the integration of life science web databases. Informatica (Slovenia), Vol. 27(1):31-4, 2003

5. Bernstein P., Haas M. Information Integration in the Enterprise. Review Articles, Communication of the ACM, Vol. 51 No. 9, 2008

6. Bhattacharjee A., Islam A., Amin M., Hossain S., Hosain H., Jamil H., Lipovich L. On- the-fly Integration and ad hoc Querying of life sciences database using LifeDB. LNCS, Springer Berlin / Heidelberg, Vol. 5690/2009, p. 516-575, 2009

7. Bilke A., Naumann F. Schema Matching using Duplicates, Intern. Conf. on Data Engineering, Tokyo, 2005

8. Bowers S., Ludäscher B. An Ontology-Driven Framework for Data Transformation in Sci- entific Workflows. Proc. DILS'04, 2004

9. Cure O., Jablonski S., Jochaud F., Abdul Rehman M., Volz B. Semantic data integration in the DaltOn system, Proceedings of the 2008 IEEE 24th Int. Conf. on Data Engi- neering Workshop, p. 234-241, IEEE Computer Society, 2008

10. Davidson S., Crabtree J., Brunk B., Schug J., Tannen V., Overton C., Stoeckert C. K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources. IBM Systems Journal, 40(2), 512-531, 2001

11. Goble C., Stevens R., Ng G., Bechhofer S., Paton N., Baker P., Peim M., Brass A. Transparent access to multiple bioinformatics information sources, IBM Systems Journal, Vol. 40 No. 2, 2001

12. Hernand T., Kambhampati S. Integration of Biological Sources: Current Systems and Chal- lenges Ahead, ACM SIGMOD Record, Volume 33 , Issue 3, 2004

13. Jablonski S., Volz B., Rehman M., Archner O., Curé O. Data Integration with the DaltOn Framework – A Case Study, SSDBM'09, Lect.Notes in Computer Science, Volume 5566/2009, p. 255-263,Springer Berlin / Heidelberg, 2009

14. Jamil H., Islam A., Hossain H. A Declarative Language and Toolkit for Scientific Workflow Implementation and Execution, International Journal of Business Process Integration and Man- agement, January 2010

15. Jones M., Ludascher B., Pennington D., Rajasekar A. Data Integration and Workflow Solu- tions for Ecology, Data integration in the life sciences DILS 2005, San Diego, CA, USA, Springer Verlag, 2005

16. Lewis S., Serale S., Harris N., Gibson M., Lyer V., Ruchter J., Wiel C., Bayraktaroglir, L., Birney E., Crosby M., Kaminker J., Matthews B., Prochnik S., Smithy C., Tupy J., Rubin G., Misra S., Mungall C., Clamp M., Apollo. A sequence annotation editor. Genome Biol 2002, 3(12). Epub 2002

17. Rahm, E. Bernstein, P. A survey of approaches to automatic schema matching,VLDB J-l,vol.10,n-4,pp. 334–350, 2001

18. Shah S., He D., Sawkins J., Druce J. Quon G., Lett D., Zheng G., Xu T., Ouellette B. Pegasys: software for executing and integrating analyses of biological sequences, BMC bioinformatics, Vol. 5 No. 1,2004

19. SnapLogic: Software solutions to your data integration challenges (home page), www.snaplogic.com, [5.2011]

20. Visser U., Stuckenschmidt H., Wache H., Vögele T. Using environmental information effi-ciently: Sharing data and knowledge from heterogeneous sources. in Environmental Information Systems in Industry and Public Administration, pp. 41–73, IDEA Group, Hershey, 2001

21. Wu J., Zhou Y. Towards Integrated and Efficient Scientific Sensor Data Processing: A Da- tabase Approach, in proceeding ACM EDBT'09, Saint Petersburg, Russia, 2009

22. Zhang Y., Boncz P. XRPC: interoperable and efficient distributed xquery. In VLDB, pages 99-110, 2007

მეცნიერულ მონაცემთა მენეჯმენტი: მიმოხილვა და კვლევის ინსტრუქციები

აბდულ მ. რეჰმანი
ბიზნესის ადმინისტრაციის ინსტიტუტი, პაკისტანი

რეზიუმე

დღევანდელი ელექტრონული სამყარო სამეცნიერო მონაცემების სამართავად, განსაკუთრებით საბუნებისმეტყველო მეცნიერების დარგში, სულ უფრო და უფრო რთული ხდება, რადგან სამეცნიერო ექსპერიმენტებისგან მიღებული მონაცემები საკმაოდ არაერთგვაროვანია მათი სტრუქტურისა (სქემა) და სემანტიკის (ინტერპრეტაცია) თვალსაზრისით. გარდა ამისა, ისეთი სამეცნიერო ინსტრუმენტების გამოყენების სიჭარბე, როგორიცაა სენსორები და მანქანა-დანადგარები, რომლებიც აგენერირებენ კომპლექსურ მონაცემებს, ართულებენ სამეცნიერო მონაცემების მართვას. ამ ტიპის მონაცემთა მართვის პრობლემების გადაწყვეტის საკითხი წარმოიშვა წინა წლებში, გაჩნდა სხვადასხვა სამეცნიერო სფეროს მრავალი მონაცემის მართვის/ინტეგრაციის სისტემები. თუმცა, ეს სისტემები არ უზრუნველყოფენ პრობლემური საკითხების იდეალურ გადაწყვეტას, უფრო მეტიც, სხვადასხვა სფეროს განსხვავებული სპეციფიკაციების, სტანდარტებისა და პლატფორმების დანართების სიმრავლის გამო, ყოველთვის არ არის იოლი შესაფერისი სისტემის შერჩევა სხვადასხვა დავალებების კონკრეტულ მონაცემთა მართვისთვის. ამ საერთო პრობლემის გადაჭრის თვალსაზრისით, სტატიაში წარმოდგენილია სამეცნიერო სფეროს მონაცემთა ინტეგრაციის არსებული სისტემების მოკლე მიმოხილვა და შემოთავაზებულია ამ მიმართულების მონაცემთა მართვის სისტემების შემდგომი განვითარების საკვანძო გზები.

МЕНЕДЖМЕНТ НАУЧНЫХ ДАННЫХ: ОБЗОР И ИССЛЕДОВАТЕЛЬСКИЕ УКАЗАНИЯ

Рееман М. Абдул
Инс-т Администрации Бизнесса, Пакистан

Резюме

В современном электронном мире управление научными данными, особенно в области естественных наук, становится все более и более сложным для домена ученых, поскольку данные, вытекающие из научных экспериментов крайне неоднородны с точки зрения структуры (схемы) и семантики (интерпретация). Кроме того, вовлечение все большего числа научных инструментов, таких как датчики и машины делает управление научными данными еще более сложной задачей, поскольку данные, полученные от таких типов инструментов, являются очень комплексными. Проблемы решения вопросов управления данными возникли в прошлые годы, были созданы многие системы управления данными и их интеграции для различных научных областей. Однако эти системы не обеспечивают идеального решения многих из существующих вопросов. Более того, в связи с высоким уровнем разнообразия в спецификациях, стандартах и из-за различных приложений платформ, выбор подходящей системы для управления конкретными данными различных задач не всегда легкий процесс. В качестве решения этой общей проблемы, в данной статье дается краткий обзор существующих систем интеграции данных научной области и предлагается ряд ключевых предложений в направлении дальнейшего развития систем управления данными.