Report on the EDBT'02 Panel on Scientific Data Integration

Omar Boucelma
Université de Provence – LSIS

Silvana Castano
Università di Milano

Carole Goble
University of Manchester

Vanja Josifovski
IBM Almaden

Zoé Lacroix
Arizona State University

Bertram Ludäscher
San Diego Supercomputer Center

1 Introduction

As scientific research becomes an increasingly larger portion of corporate expenditures, pressure is mounting to make the processes more efficient. Data acquisition, access, management, analysis and the sharing of all available resources will be at the core of the transformations needed to achieve the next level of efficiency in research organizations. However current data management technology is geared toward business data and several technical challenges remain to make it suitable for scientific data. A panel entitled *Scientific Data Integration* was held on March 25^{th}, 2002, at the 8^{th} Conference on Extending Database Technology (EDBT) in Prague, in the Czech Republic. The panel focused on the issues needed to be addressed to enable scientific data integration and discussed solutions. Omar Boucelma, Université de Provence, and Zoé Lacroix, Arizona State University, moderated the panel which included Silvana Castano, Università di Milano, Carole Goble, University of Manchester, and Bertram Ludäscher, San Diego Supercomputer Center.

2 Issues in Scientific Data Integration

Scientific data integration first raises traditional issues of data integration. Silvana Castano presented an overview of these issues (see Section 2.1). The specificity and complexity of scientific data significantly worsen the problem. Issues related to the integration of biological data were presented by Carole Goble and Bertram Ludäscher.

2.1 Data Integration

The goal of data integration is to construct a global description, called global schema, of the data coming from a multitude of heterogeneous sources. The global schema is a reconciled view of the underlying data sources and provides a uniform query interface exploited by the user to pose queries independent from source location and heterogeneity. A data integration system is a complex system, generally based on a wrapper/mediator architecture, and is responsible for creating the global schema and for processing queries over it, interfacing the underlying data sources. Silvana focused on the following key issues for scientific data integration, based on her experience in the framework of the ARTEMIS/MOMIS data integration project [BCVB01].

2.1.1 Information extraction and metadata

Data integration systems generally follow a "semantic approach" to integration based on the conceptual schemas or metadata of the sources to be integrated and on a middleware data model for a uniform and semantically rich representation of heterogeneous sources (e.g., structured, semistructured, or unstructured). Wrapper tools are used to translate the source data models into the middleware data model, which is used also to represent the global schema. To enable scientific data integration, an important requirement is thus the availability of metadata-based infrastructures and intelligent information extraction techniques.

2.1.2 Data semantics and matching techniques

A key issue is the capability to bring data semantics aspects in automated schema matching techniques for semantic data integration. To this purpose, hybrid schema matching techniques considering schema names, descriptions, data types, relationship types, constraints and structure of a source schema are essential to recognize portions of different

source schemas that denote the same concept and define semantic mappings between them. Matching techniques take into account extensional interschema knowledge are also required, to ensure a correct reconstruction of instances spanning multiple sources [BCVB01]. Instance-based matchers considering data instances can provide relevant information when schema information is limited or poor. Development of generic and robust mixed matching techniques, providing support for using multiple kinds of knowledge on data semantics together with the reuse of known matches are important directions to be explored to cope with integration requirements posed by scientific data.

2.1.3 Generation of a Global Integrated Representation

The global schema provides an appropriate abstraction of all the data residing at the sources. An important aspect is the definition of mappings that specify reciprocal correspondences between the global schema and the source schemas and the choice of the most appropriate architecture for populating global schema (i.e., virtual or materialized or a combination of the two), taking into account huge volumes of source data and maintenance costs. Another aspect to be considered is related to the generality of the global schema. A single schema for scientific data integration is clearly unfeasible, given the huge number of involved data sources. A scenario where several global schemas co-exist for different fields/communities is expected, and query processing techniques operating under an open-world assumption are required.

2.2 Issues Specific to Scientific Data Integration

Carole set the scene for integration of scientific data by referring to her experiences in integrating information in Life Sciences. She pointed out that experimentation in biology is now as much performed in-silico, over large complex, distributed experimental data sets, as it is at the bench. Each area of molecular biology generates their own data repositories in many representation forms (literature, images, video, free text annotations), many of which are not immediately computationally accessible. The various repositories have different data formats, access mechanisms and user interfaces and have different terminologies too. A plethora of specialist interrogation and analysis tools exist, each typically associated with a particular database format. Around 400 public databases are currently published and used, ranging from reference database curated by government labs to "mom and pop" specialist repositories.

Carole showed that although the resources have usually been developed independently, the complex questions and analysis posed by biologists cross the artificial boundaries set by these data banks. Hence, the key task was one of integration and fusion, linking together pieces of evidence from different information sources, different experiments, different workers; comparing results, inferring and testing new hypotheses and generating new insights. This problem was likely to get worse as the variety of information that must be combined is expanding (e.g. genomic, proteomic, transcriptomic, etc.) and related in complex and ill-understood ways (e.g. metabolic pathways).

As this is such a pressing problem, a large number of proposals have been put forward, and used, for bioinformatics resource integration systems. Carole listed a number of solutions. Conformance to a standard terminology is a popular and effective mechanism for integrating content by mapping to it a standard vocabulary. For example, the Gene Ontology, a classification and vocabulary for gene function, is used by 14 major data resources, facilitating the use of domain maps (see Section 3.1) as means of linking instances. Tightly coupled middleware developed on the mediator/wrapper paradigms for both federated integration and data warehouses has also been extensively explored. She presented an impressive list, including: TAMBIS [GSN+01], IBM's DiscoveryLink, GeneticXchange's K1, BioKleisli, ISYS, OpenMMS, OPM, and Lion BiosSciences' SRS. She also described recent developments using loosely coupled open integration middleware solutions based on the publication of service descriptions and protocols, arising from the Grid computing initiative [FKT 01] (e.g. myGrid, Obigrid, ProteomeGrid) (`http://www.gridforum.org`); OMG Life Science Research (`http://www.omg.org/homepages/lsr/`) the I3C vendor initiative (`http://www.i3c.org/`) and the Open Bio Foundation bioMOBY project (`http://www.biomoby.org/`). All these lean on Web Services and XML-based solutions.

Carole posed the question that, given this activity, what are the challenges? She proposed three (cf. Section 3 which describes new approaches for and experiences with integration models that are more adequate for scientific data):

1. The nature of the data itself. The data is highly heterogeneous and diverse, with different formats, structure, schemas, and coverage. The databases and tools are highly autonomous and regularly change, as the data and the understanding of how to describe the data, changes.

2. The way the data is published and managed,

which assumes that humans will read the entries and integrate by navigation through connected instances. Although the number of databases stored as databases (i.e., not flat files) is increasing, this does not mean that the repositories have open SQL interfaces. Most interfaces are point and click, without APIs, without a query interface and often without a published schema. This meant that screen scraping results is de rigueur. As the databases are rarely interacted with by SQL, but rather by call interfaces, there is no distinction made between a database and an analytical tool. Integration is popularly through "click navigation", with integration pivoting about models and instances that biologists understand rather than schema-based. For example, DAS (`http://www.biodas.org/`) uses a protein sequence as an integration model. Integration is seen by biologists as an interactive workflow, piping the outputs of one tool or database into the input of the next. Carole also raised the added complication that biology is a descriptive as well as a numeric discipline, so much of the data is record in narrative "annotations" associated with primary data. Hence, "schema-less descriptions" were widespread; the data repository being self-described through mark-up in order to cope with change. Thus there was a great emphasis on controlled vocabularies and integration through content independent of matching schematic components.

Carole pointed out that curating a data entry's annotation is a well-established means of pooling and publishing information. The so-called "annotation pipeline" is one where primary raw results are entered into a database; the data analyzed and annotated with new knowledge (usually manually) and placed in a secondary database. Further analysis and integration lead to data being drawn from a number of databases into another's data entry. The annotations build in automated cross-referencing to other data instances in other databases (independently owned). She suggested that annotations could be thought of as the recording of intellectually handcrafted derived information borne of integration. The effect is that annotations are propagated and integrated down a chain of databases.

3. The nature of scientific endeavor is such that fluid and dynamic federation are formed, especially in discovery by the rapid mustering and disbanding of resources. This makes the integration problem even more one of workflow and of building virtual organizations. An annotation pipeline, for example, is a workflow, and the resources used to build the annotation is a virtual organization. Carole raised the important issues of provenance, quality and change notification – what if the data on which an annotation is derived changes? What was the process of derivation? Can I repeat or re-enact the integration? Do you trust the results? What would have been the integration result at a particular time based on hat we knew then? Science has a guardianship of knowledge. Moreover, the information is both accumulative (it is dangerous to alter anything) and open to reinterpretation.

3 Scientific Data Integration Support

The issues presented in Section 2 were addressed by the panelists in several of their projects. This section presents solutions and areas of research that aim to successfully resolve issues on scientific data integration. Bertram Ludäscher reported on experiences in scientific data integration from the Neuroscience domain (Section 3.1). Silvana Castano emphasized on the need for an ontology architecture for information integration in web-enabled systems that provides a shared understanding of data semantics (see Section 3.2). Carole Goble focused on providing transparent access to resources with an ontology (see Section 3.3) and the need to integrate scientific services (see Section 3.4).

3.1 Scientific Data Integration Needs KR+DB

Bertram started by explaining that his experiences with data integration in the Neuroscience domain have shown that current database mediation technology breaks down for some complex scenarios in which a combination of techniques from database mediators and knowledge representation is necessary [BIRN]. Here, "complex" refers to the fact that domain experts have to be involved in order to come up with meaningful mediated views in the first place. He has developed, together with his colleagues, a technique called *model-based mediation* (MBM) that has brought together databases and knowledge representation for the purpose of scientific data integration across different and complex domains [LGM01].

One goal of MBM is to turn domain scientists' *questions* into *database queries* that can be evaluated against multiple sources. For example, a neuroscientist may ask: *"What is the cerebellar distribution of rat proteins with more than 70% homology with human NCS-1? Is there any structure specificity? How about other rodents?"*. Such question can, in principle, be answered using sources that export protein localization data (PROTLOC), information on calcium binding proteins (CAPROT), morphometry data (SYNAPSE), and others [LGM00].

Bertram claimed that the primary difficulty was *not* in the legendary impedance mismatch between different data models, or schema heterogeneities, but rather that there are "semantic gaps" between the source data which need to be filled with "glue knowledge" from domain experts, in order to relate item X from one source with item Y from another source. He advocated for a common "semantic coordinate system" that provides a reference mechanism to link a source's data objects to concepts at the mediator. *Ontologies*, i.e., formal representations of a conceptualization provide such a coordinate system and are one component of a model-based mediator system. In MBM so-called *domain maps* (a.k.a. ontologies) provide this terminological glue knowledge; for example, a domain map of anatomical structures ANATOM has been used to integrate data from different species, scales, and resolutions. Thus the integration mechanism relies on data instances conformance to a shared set of terms.

Bertram went on to say that (a) sources should not simply export a database schema but also the *semantic types* of the represented entities, so that exported data can be "understood" by the mediator at the conceptual level; and (b) that sources should not only export their local object model, but relate ("*contextualize*") it by specifying its properties relative to the shared domain maps that have been registered with the mediator. He proposed approaches using rich object-oriented models and declarative rule languages (e.g., F-Logic [KLW95]) for the former, and description logics to interrelate and reason with concepts for the latter [Hor98].

3.2 Scientific Data and the Semantic Web

Silvana contended that the integration of scientific resources will be greatly improved by the current attention on the development of techniques and tools for the Semantic Web. In the literature, main research issues are concerned with the development of methods and tools for the construction of concept ontologies and the definition of thematic views to improve semantic interoperability and to effectively support users in search activities. Recent developments focused on the design of ontologies to support the integration of heterogeneous data sources. Ontologies generally provide a common vocabulary to support the sharing and reuse of knowledge. By coupling basic features of Semantic Web ontologies with those of semantic data integration system, the integration knowledge of a set of data sources in a given domain can be organized according to a three-layer Web domain ontology architecture [CDDM01]. She went on to outline the layers: (i) a *semantic mapping scheme*, representing knowledge about similarity relationships (i.e., semantic mappings) between matching elements of different data sources; (ii) a *mediation scheme*, representing interorganizational knowledge in form of global concepts, mediating between the possibly heterogeneous representations of the concept in the different data sources; and (iii) a *categorization scheme*, representing the interorganizational knowledge according to topic-based views, using hierarchically organized subject categories capturing the semantic meaning of a set of inter-related global concepts.

She went on to discuss how the three-layer modeling schemes provides a Web search space (to locate the data sources associated with a given subject category and/or global concept) and a semantic search space (where users can browse on concept networks at different layers through intra-layer semantic links, and move across layers through inter-layer semantic links).

3.3 Using Ontologies to Provide Transparency to Users

Silvana proposed that the Web domain ontology supports mediator services for query formulation and processing, to correctly answer a query issued by an user formulated on the structure of a global concept. By exploiting appropriate mapping rules, such a query can be reformulated on the specific terminology of each data source concept involved in the definition of the global concept, and then the retrieved data can be combined and expressed in terms of the global concept structure, understood by the user. Carole picked up this point to show how ontologies [BBP+99] can be used to provide scientists' a transparent access to integrated scientific resources by describing the TAMBIS project.

The Transparent Access to Multiple Bioinformatics Information Sources (TAMBIS) project citegoble-01 aimed to provide the user with maximum transparency when accessing diverse bioinformatics data sources, shielding the users from those sources to the extent that the users cannot see the individual sources. It provides the illusion of a single query language, a single data model and a single data location. TAMBIS achieved this by using: a single terminology based on an ontology of molecular biology and bioinformatics [BBP+99] against which the user can formulate queries; and mappings from terms in the conceptual representation onto terms in external sources. The sources were wrapped to give the illusion of a common request language for each information resource. TAMBIS therefore provided a level of indirection between the user and the external

sources. By implementing the TAMBIS system Carole came across all the difficulties that were identified in Section 2.2: none of the sources used were published as relational databases with SQL query capability so a biologist could ask smarter questions than the databases would permit to be answered; much of the information was in the annotations and inaccessible to retrieval; the high source autonomy lead to serious wrapper fragility and the need for semi-automated source management to dealing with new and changing sources. Moreover, returning results drawn from a range of data sources wasn't enough – the results needed to be adorned with their provenance in order to give some rudimentary support for data quality. The "T" in TAMBIS was to hide the resources from the user in the traditional way of federated databases; in fact the users did not want transparency, they wanted to interfere in the choice of repositories, wanted to inspect intermediate results, and demanded explanations for query plans, and the right to alter them. Carole said that she realized that, although they had set out to build a traditional federated virtual database, they had actually built a workflow system. This was because many of the resources were analytical functions, so query plans were constrained by appropriate ordering to string together processes. As expected, the global schema needed considerable maintenance, and this was a cost.

3.4 Integration of Scientific Services

Following on from Caroles workflow insight, the panel discussed the fact that in addition to model-based mediation which can be seen as a combination of traditional mediation technology (global-as-view, view unfolding, query rewriting and optimization) with knowledge representation techniques (source registration relative to shared ontologies/domain maps, semantic integrity constraints), there is an increased interest in *scientific workflow management*. Scientists not only glue together data sources but also application programs such as analysis and modeling tools and simulation packages [SDM, SEEK]. Unlike in the above mentioned scenarios, in this setting, scientists chain together remote database queries and calls to application programs in a more low-level and procedural manner; in particular, the sequence of steps in a chain of operations is crucial (whereas a database mediator may reorganize the initial unfolded query plan more freely).

Scientific services may also be delivered and integrated through a large scale distributed platform known as *Grid computing*. The purpose of the Grid is to deliver a collaborative and supportive environment that allows geographically distributed scientists to achieve research goals more effectively. The Grid aims to provide a persistent, high performance, reliable, high capacity set of middleware services that can be rapidly scaled and readily managed. The Grid development includes distributed storage and computation, data access and analysis, certificate authorities and policies, protocols for discovery and access, peer-to-peer computing, etc.

Many Grids are being developed: myGrid, DOE Science Grid, EuroGrid, TeraGrid, etc.[1] myGrid aims to design, develop and demonstrate higher level functionalities over an existing Grid infrastructure that support scientists in making use of complex distributed resources. The myGrid project follows the more loosely coupled model of the Grid and Semantic Web Services. The Grid is explicitly about middleware and services to build dynamic virtual organizations, with support for security, ownership, transactions, state management etc (`http://www.gridforum.org`). Semantic Web Services describe and discover services using the techniques for describing and reasoning about metadata drawn from the Semantic Web (`http://www.semanticweb.org`). myGrid (`http://www.mygrid.org.uk`) is a middleware for a platform to support personalized data intensive in silico experiments for biologists. The central interoperation mechanism is workflow, and provenance management is to the fore. There is still a role for traditional virtual data integration, but as part of the solution, not the whole solution. The panel wondered if the Grid approach more closely reflected the nature and requirements of scientific information integration.

4 Conclusion

Scientific data integration offers new exciting areas of research for the database community. The need for systems integrating complex, dramatically heterogeneous scientific resources will only increase in the future. However, many assumptions of traditional database development need to be revised in order to build scientific integration systems. The panel has identified issues that challenge traditional assumptions. Scientific data organization change constantly over time, therefore an approach based upon a semi-structured data representation is preferable to a traditional pre-defined database schema. Scientific data provides many levels of semantics: the use of ontologies enables a semantic integration of scientific data. Scientific data integration also requires the integration of the

[1] For a list of Grid initiatives see `http://www.gridforum.org/L_Involved_Mktg/inint.htm`.

technology to manipulate, analyze, and visualize them. To provide scientists with the integration of a variety of scientific services, an approach such as Grid computing seems to be more appropriate.

Finally, although the panel was centered around biological data integration, issues and solutions discussed showed many similarities with other scientific domains such as geosciences and ecology [GEON, ECOINF].

References

[BBP+99] P. Baker, C. Goble S. Bechhofer, N. Paton, R. Stevens, and A. Brass. An Ontology for Bioinformatics Applications. *Bioinformatics*, 15(9):510–520, 1999.

[BCVB01] S. Bergamaschi, S. Castano, M. Vincini, and D. Beneventano. Semantic Integration of Heterogeneous Information Sources. *Data and Knowledge Engineering*, 36(3), 2001.

[BIRN] Biomedical Informatics Research Network Coordinating Center (BIRN-CC), U.C. San Diego. `http://nbirn.net/`, 2001.

[CDDM01] S. Castano, V. De Antonellis, S. De Capitani di Vimercati, and M. Melchiori Designing a Three-Layer Ontology in a Web-based Interconnection Scenario. In *Proc. of DEXA WEBH Workshop*, September 2001.

[ECOINF] Ecoinformatics – An online data and information management resource for ecologists. `http://www.ecoinformatics.org/`, 2002.

[FKT 01] I. Foster and C. Kesselman and S. Tuecke. The anatomy of the Grid: Enabling scalable virtual organizations. *The International Journal of Supercomputer Applications*, 2001.

[GEON] Geoinformatics Network. `http://www.geoinformaticsnetwork.org/`.

[GSN+01] C. Goble, R. Stevens, G. Ng, S. Bechhofer, N. Paton, P. Baker, M. Peim, and A. Brass. Transparent access to multiple bioinformatics information sources. *IBM Systems Journal*, 40(2):532–551, 2001.

[Hor98] I. Horrocks. Using an Expressive Description Logic: FaCT or Fiction? In *Principles of Knowledge Representation & Reasoning*, 1998.

[KLW95] M. Kifer, G. Lausen, and J. Wu. Logical Foundations of Object-Oriented and Frame-Based Languages. *Journal of the ACM*, 42(4):741–843, July 1995.

[LGM00] B. Ludäscher, A. Gupta, and M. E. Martone. Model-Based Information Integration in a Neuroscience Mediator System. In *VLDB*, Cairo, 2000. System demo.

[LGM01] B. Ludäscher, A. Gupta, and M. E. Martone. Model-Based Mediation with Domain Maps. In *17th Intl. Conf. on Data Engineering (ICDE)*, Heidelberg, 2001.

[SEEK] Analytical Pipelines for the Science Environment for Ecological Knowledge (SEEK). `http://kbi.sdsc.edu/AP/`.

[SDM] Scientific Data Management Center (SDM). `http://sdm.lbl.gov/sdmcenter/`.