SciPort: An Adaptable Scientific Data Integration Platform for Collaborative Scientific Research

Fusheng Wang
Integrated Data Systems Dept
Siemens Corporate Research, USA
fusheng.wang@siemens.com

Pierre-Emmanuel Bourgué*
Université de Technologie de Compiègne, France
pbourgue@gmail.com

Georg Hackenberg*
University of Mannheim, Germany
ghackenberg@gmail.com

Mo Wang*
University of Duisburg-Essen, Germany
mo.hy.wang@gmail.com

David Kaltschmidt*
Freie Universität Berlin, Germany
david@inf.fu-berlin.de

Peiya Liu
Siemens Corporate Research, USA
peiya.liu@siemens.com

ABSTRACT

Scientific data are posing new challenges to data management due to the large volume, complexity and heterogeneity of the data. Meanwhile, scientific collaboration becomes increasingly important, which relies on integrating and sharing data from distributed institutions. In this demo, we present SciPort, a Web-based platform on supporting scientific data management and integration based on peer-to-peer architectures, where researchers can easily collect, publish, and share their complex scientific data across multi-institutions. SciPort provides a general metadata based data model to capture the context description of experiments and link experiment data into comprehensive metadata documents, and supports a hierarchical organization of the overall data space for data browsing. SciPort takes two alternative "peer"-to-"peer" (or peer-database-to-peer-database) based approaches to integrate scientific data: pure peer-to-peer architecture and central server based peer-to-peer architecture. The later provides a virtual view of all published data from multiple local sites and supports complex queries with XQuery. The system provides a unified framework for adaptable architectures and customizable schemas, and supplies comprehensive tool set to manage and share scientific data. SciPort was first prototyped in Siemens Corporate Research, and now becomes a mature product and has been successfully used in both biomedical research and clinical trials for scientific research communities.

1. INTRODUCTION

Scientific research is increasingly relying on collaborative effort across multiple research groups and inter-disciplinary consortium. For example, NIH provides large-scale collaborative project awards for a team of independently funded investigators to synergize and integrate their efforts. Indeed, the awards require the research results to be shared. One example is the "Networks for Translational Research: Optical Imaging" (NTROI), research consortia consisting of more than 20 universities, in which there are hundreds of researchers located in distributed institutions working on the problems of optical imaging. One essential need for such consortia is to promote data sharing across multi-organizations.

As another example, Siemens Medical Solutions has research collaboration with hundreds of research sites distributed across the US, each providing Siemens marketing support by periodically delivering white papers, case reports, clinic methods, application tips, clinic protocols, state-of-the-art images, flash articles, etc. In the past, data were delivered through emails, CDs, hard copies, etc. As a result, deliverable content was non-centralized, and was difficult to manage, integrate, search, and reliably archive.

Besides, clinical trials are often distributed among multiple hospitals or medical research institutes. For example, University of Toronto TiBS Center is performing clinical trials among several hospitals, and University of California, Irvine is leading a group of universities to conduct Chemotheraphy based clinical trials. Patients are recruited at distributed institutions and experiments are performed on these patients. These require a platform to collect both clinical data and experiment data at multiple distributed institutions, and integrate them together for patient study and data analysis.

While there is a strong demand of managing and sharing scientific data, many challenges exist. Scientific data have high complexity and diversity, and are often in large scales. New technology advance brings diverse instruments and dynamic computation tools, leading to heterogenous data formats such as medical images, spreadsheets, PDF files, XML documents, and many others. Especially, for biomedical research, the majority of the data is stored as images and files, and the data size can be extremely large. The mix of all types of scientific data demands an adaptable platform that can provide general data modeling and management of scientific experiments.

While the need for collaboration and integration keeps on increasing, scientific data tend to be isolated. Indeed, re-

*Work done while visiting Siemens Corporate Research

VLDB '07, September 23-28, 2007, Vienna, Austria.

searchers would rather have the best control of their data on a server located on their own labs, instead of "outsourcing" them somewhere else. Each researcher and its collaborators will naturally form a unit of data source of themselves.

These new trends demand a general experiment management platform for collaborative scientific research, which can provide: i) a general data model to represent scientific experiments, so researchers can easily represent and organize their data and experiments; ii) context description of experiments, thus experiments can be understood, repeated and shared easily; iii) proper classification and placement of experiments in the collaborative community, thus experiments can be classified and identified and browsed; iv) convenient tools to help data providers to generate such context information; and finally, iv) a system architecture that provides transparent integration of data and experiments across all institutions for sharing and searching.

These requirements however cannot be supported by past work [1, 2, 3]. Traditional file-based P2P networks can support sharing of files based on very limited metadata, but difficult to support management and sharing of complex scientific experimental data. Scientific data management products such as [4] have no support of data sharing across multiple distributed institutions. Grid-based scientific integration such as BIRN [5] needs sophisticated deployment and is often used to connect small numbers of sites, and doesn't address directly the issues on scientific data collecting and management.

SciPort: an Adaptable Platform for Scientific Data Management and Integration

We develop SciPort, a Peer-to-Peer based platform to integrate scientific data from multiple data sources located at distributed locations, to promote collaborative scientific research. Here we extend the concept of peer-to-peer as database-to-database, with a mix of structural metadata and files/images.

Our contribution is summarized as follows.

- SciPort provides a unified framework to model complex scientific experiments by capturing all the context information of experiments, and uses a metadata based approach to represent all context information - including metadata fields and files;
- SciPort is high adaptable, thus users can easily customize their own applications without requiring the expensive and time-consuming services of database administrators or programmers;
- SciPort uses flexible data integration and sharing architectures. SciPort provides two alternative architectures to integrate data from multiple sites, pure P2P based integration for loose collaboration, and central server based P2P integration for close collaboration. The later works like a "Napster" for distributed scientific databases. The architecture is very flexible, and local sites can easily join in.

In this demo, we will show how a Local SciPort Server manages scientific data, and how the integration architectures work. Especially, we provide solutions for several challenging issues in our data integration architectures, which were not addressed in [6, 7].

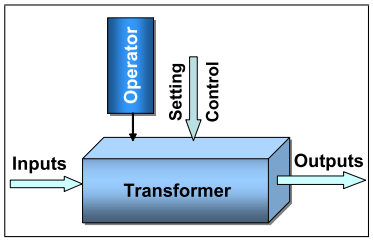

Figure 1: Transformation-based Modeling of Scientific Experiments

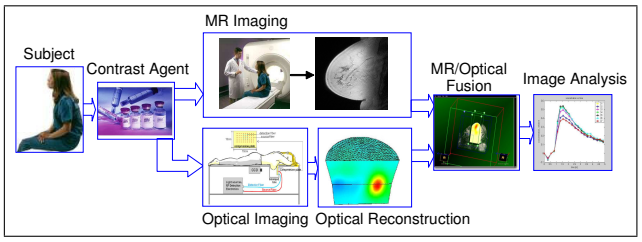

Figure 2: A Sample Experiment of MR/Optical Fusion

2. SCIPORT FOR DATA MANAGEMENT

Each Local SciPort Server serves as a scientific data management system independently. We provide a general data model that can model complex scientific data, and we use metadata to describe all context information of an experiment.

Scientific experiments are characterized by several essential properties: i) *Transformations*: steps or processes to perform experiments; ii) *Lineage*: processing history of a data product, i.e., how transformations are related and linked together; and iii) *Classification*: correctly placing an experiment with respect to the overall data space in the research hierarchy. We develop the following approaches to model and manage scientific experiments.

Transformation-based Data Modeling. We define *transformation* (Figure 1) as the fundamental object in scientific experiments, which encodes an element of an experimental input/output process in enough details that the process could be repeated by others "skilled in the art." A transformation consists of a transformer, input, output, operator, and the setting. Individual "atomic" level transformations can be linked into larger transformation pipelines. For example, the MR/Optical fusion experiment in Figure 2 can be viewed as the following transformations: contrast agent, MR imaging, optical imaging, optical reconstruction, MR/optical fusion, and image analysis.

Hierarchical Data Organization. This provides a vertical organization of experimental data across multiple institutions. Indeed, a research consortium can be naturally modeled as a tree structured hierarchy, thus data and experiments can be quickly browsed and identified through this hierarchy. For example, the following is a sample hierarchy for Siemens Medical Research Collaboration:

```
Research Collaboration -> Modality -> Research Site ->
  Group -> Deliverables
```

Metadata Management. We use metadata to describe experiments and their transformations, and link experiment data together through metadata. Metadata of experiments are represented as XML documents, thus can be easily indexed and searched through standard XML query languages. The metadata document together with the linked data files preserves all the information of a transformation. To provide maximum flexibility, the schemas of metadata can be customized by describing them in a unified interface using XML.

SciPort provides a Web-based semiautomatic process for researchers to collect metadata. Authored metadata documents together with their linked data files are stored in SciPort Local Servers. The metadata XML document will then be automatically indexed into the index server for efficient retrieval.

3. SCIPORT FOR SCIENTIFIC DATA INTEGRATION

Based on the level of integration and collaboration, users can choose from two alternative integration architectures. For loosely coupled collaboration, users can manage their own data independently at their Local SciPort Servers, and share data through a pure P2P approach; and for closely coupled collaboration, where the integrated organization and sharing of data is important, a SciPort Central Server is used to provide an integrated view of all shared data, and support collaboration among multiple sites.

3.1 Pure Peer to Peer Based Data Integration Architecture

The pure P2P based data integration architecture is illustrated in Figure 3. For this architecture, data are shared through queries: a query on metadata is sent from a Local Server and spread out to all other Local Servers defined on a control list, each Local Server checks access permissions and processes the remote queries, and the list of results is then aggregated and sent back to users. Data then can be downloaded from peer sites from the links in the result. In this approach, schemas and global data organization are unknown, and data are shared through probing other peer sites. The disadvantage is that due to the loose collaboration, it is hard to share schemas or to get an overall view of all data: data are viewable only through probing with queries. Moreover, security is a concern for this architecture, and has to be enforced.

To enforce security, a Local Server defines two lists: the list of Local Servers that can access this Local Server, and the list of Local servers that this Local Server will query upon. To enhance data protection, we enforce a public key/private key pair to be exchanged between two Local Servers prior to connections.

3.2 Central Server Based Peer-to-Peer Integration Architecture

To support closer collaboration and integration, we develop a Central Server based P2P architecture to share and integrate data (Figure 4). This architecture uses a Central Server to keep track of all data stored at peers (Local Servers), and peers are responsible for hosting the data. Each Local Server can work independently as a server for

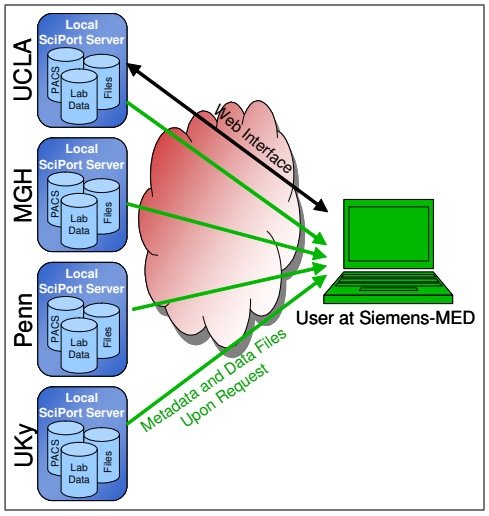

Figure 3: Pure P2P Architecture for Scientific Data Integration

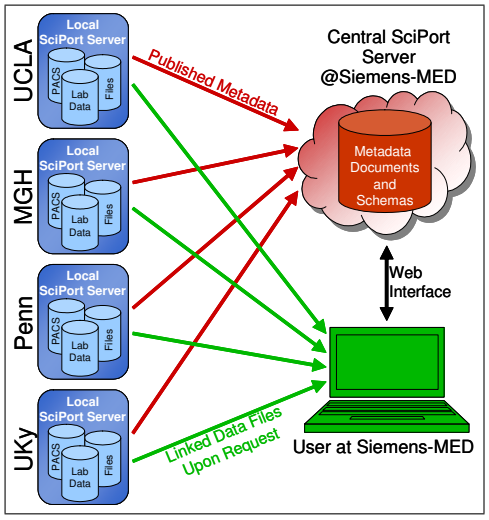

Figure 4: Central Server Based P2P Architecture

managing scientific data for a local institution or project. After a local user authors a document, he/she can selectively publish the document to the Central Server, while the original scientific data files are still stored at the Local SciPort Server. The Central Server is first a directory server of experiments. It keeps a directory of all scientific data and the transformation schemas by indexing all metadata XML documents, and maintains a global hierarchical organization of all instances across multiple Local Servers, with links to their storage locations (URL) at local servers. This central directory architecture provides an integrated view of experiments across all collaborative sites.

One challenging issue for the central server based P2P architecture is synchronizing hierarchy and schemas between Local Servers and the Central Server. We support several scenarios: i) Central Server controlled hierarchy and schemas. Administrators on the Central Server enforces the data organization hierarchy at Local Servers and the transformation schemas to be used for each group. For ex-

ample, Siemens Medical Solutions takes this approach to manage collaboration with hundreds of research sites; ii) Local Server controlled hierarchy and schemas. Hierarchy and schemas are defined autonomously and published to the Central Server. This scenario applies to cases that different research sites run relatively independently; iii) Collaboratively managed hierarchy and schemas. The Central Server can control high level hierarchy and enforce it to Local Servers, and the Local Servers control lower level hierarchy such as experiments and studies. For schemas, Local Servers can collaboratively authoring and defining schemas through the coordination of the Central Server. Schema sharing and versioning is also possible through the Central Server.

4. DEMONSTRATION

In this demo, we will demonstrate the SciPort system with its major functionalities and tools, with real data from our customers, including Siemens Medical Solutions Research Collaboration, and NTROI research consortium. SciPort is built with J2EE and XML, running on Apache Tomcat servers and BerkeleyDB XML database server. The system is OS neutral and can run on any machines. The data sizes from our customers range from hundreds of metabytes to terabytes.

We will show the architecture and workflow of SciPort for two scenarios – the loose integration scenario and the close integration scenario, and how they are used for our customers. We will demonstrate how to easily adapt the system to different applications through schema management and hierarchy setup, and how to author, share, integrate, browse, search and aggregate scientific data.

The schema management tool supports comprehensive model and design of scientific experiments, ranging from scientific data to clinical data. The tool also provides schema sharing across multiple sites through the Central Server. This tool allows users to configure their own applications without requiring the expensive and time-consuming services of database administrators or programmers.

The hierarchy setup tool provides GUI to design and setup hierarchies for research collaboration groups, and assign schemas into the hierarchy (at arbitrary node). The hierarchy browsing tool provided by SciPort makes it very convenient to browse documents in a hierarchical way.

One major tool provided by SciPort is easy data collection through its comprehensive authoring tool. An authoring form is automatically generated through the schemas defined by users when the schema is selected through a hierarchy. The tool can be used to collect arbitrary data, including textual data, image data, files, and even DICOM images directly from PACS servers. This makes it extremely convenient for collecting complex scientific data. The authoring tool also provides fine-grained access at document level, making the documents highly secured.

The publishing tool makes it possible for users to selectively publish their metadata documents into Central Server(s), while gaining full control of their original data.

The query and aggregation tool provides maximum flexibility for users to perform keyword and regular expression based search, structural search, or combination of these, as well as joins on multiple schemas, result filtering and aggregation. The queries are implemented as XQuery on XML databases.

More information on SciPort software can be found at SciPort Wiki [8].

5. CONCLUSION

In this demo we present SciPort, a Web-based platform for scientific data management, integration and collaboration, to support collaborative research effort across multi-institutions and interdisciplinary teams. SciPort first provides a scientific data management system that can run independently at individual research sites by providing data collecting, modeling, and searching. By taking these Local SciPort Servers as peer nodes, we can now integrate scientific data in two alternative peer-to-peer based architectures: i) a pure peer-to-peer architecture where searching is distributed across multiple peer nodes, and ii) a central server based peer-to-peer architecture where metadata are published and shared on a central server; the Central Server provides a rich directory for indexing and searching across all research sites.

By providing a unified and effective means for data modeling and access, SciPort provides a flexible and powerful platform for sharing scientific data for scientific research communities, and has been successfully used in both biomedical research and clinical trials, such as UCI, UPenn, SUNY Downstate, UToronto, Dartmouth, MGH, UCLA, Siemens Medical Solutions, etc.

6. ADDITIONAL AUTHORS

Additional authors: Cornelius Rabsch*(University of Mannheim, email: rabsch@db.informatik.uni-mannheim.de), Patrick Kling*(University of Waterloo, Canada, email: pkling@cs.uwaterloo.ca), Gerald Madlmayr*(Johannes Kepler University Linz, Austria, email: Gerald.Madlmayr @ fh-hagenberg.at), John Pearson (Siemens Corporate Research, USA, email: pearson.john@siemens.com) and Joe Carpinelli (Siemens Corporate Research, USA, email: joe.carpinelli@siemens.com)

7. REFERENCES

[1] W. S. Ng et al. PeerDB: A P2P-based System for Data Sharing. In *ICDE*, 2003.

[2] C. Li el al. RACCOON: A Peer-Based System for Data Integration and Sharing. In *ICDE*, 2004.

[3] The Hyperion Project. http://www.cs.toronto.edu/db/hyperion/.

[4] Axiope Catalyzer. http://www.axiope.com.

[5] Biomedical Informatics Research Network. http://www.nbirn.net/.

[6] F. Wang, P. Liu, J. Pearson, F. Azar, and G. Madlmayr. Experiment Management with Metadata-based Integration for Collaborative Scientific Research. In *ICDE*, 2006.

[7] F. Wang, C. Rabsch, P. Kling, P. Liu, and J. Pearson. Web-based Collaborative Information Integration for Scientific Research. In *ICDE*, 2007.

[8] SciPort Wiki. https://sciportserver.scr.siemens.com/mediawiki.