User Centred Networked Health Care
A. Moen et al. (Eds.)
IOS Press, 2011

doi:10.3233/978-1-60750-806-9-867

Service Oriented Data Integration for a Biomedical Research Network

Matthias GANZINGER[a,1], Tino NOACK[a], Sven DIEDERICHS[b,c], Thomas LONGERICH[c], Petra KNAUP[a]
[a] *Department of Medical Informatics, University of Heidelberg.*
[b] *Helmholtz-University-Group "Molecular RNA Biology & Cancer", German Cancer Research Center (DKFZ).*
[c] *Institute of Pathology, University of Heidelberg.*
Heidelberg, Germany.

Abstract. In biomedical research, a variety of data like clinical, genetic, expression of coding or non-coding ribonucleic acid (RNA) transcripts, or proteomic data are processed to gain new insights into diseases and therapies. In transregional research networks, geographically distributed projects work on comparable research questions with data from different resources and in different formats. Providing an information platform that integrates the data of the projects can enable cross-project analysis and provides an overview of available data and resources (tissue, blood, etc.). For a German liver cancer research network consisting of 22 individual projects, we develop the integrated information platform pelican – **p**latform **e**nhancing **li**ver **ca**ncer **n**etworked research. In our generic approach, data are made available to the research network by standardized data services based on technologies provided by the cancer Biomedical Informatics Grid (caBIG). It has shown that publishing service metadata in a corresponding repository is a major prerequisite for automated discovery, integration, and conversion of data records and data services. We identified data confidentiality and intellectual property considerations as major challenges while establishing such an integrated information platform. As a first result we implemented a working prototype to validate our approach.

Keywords. biomedical research, service oriented architecture, data integration

1. Introduction

Biomedical informatics research can provide resources that represent, visualize and analyze large-scale genetic data efficiently and flexibly [1]. Nevertheless, a lack of interoperability among data resources from independent institutions is described as a severe problem for biomedical research in current literature [2]. The variety of representations and semantics usually leads to data sets that are stored in heterogeneous formats, described with different terminologies and analyzed with dedicated applications. This heterogeneity may hamper the development of new strategies targeting cancer [3] and their translation from bench to bedside. Several approaches have been started to address this problem. Data warehouses are introduced, so that data

[1] Corresponding author: Matthias Ganzinger, Dpt. of Medical Informatics, University of Heidelberg, Im Neuenheimer Feld 305, 69120 Heidelberg, Germany; E-mail: matthias.ganzinger@med.uni-heidelberg.de

from biological databases can be integrated, locally stored, and analyzed [4, 5]. It has been recognized that collecting information from different areas of research offers important advantages: Relevant independent results are tied together and specialists are pointed into new directions [6].

In Germany, a transregional research network (TRN) on hepatocellular carcinoma (HCC) has been established. Within the TRN, 22 biomedical research projects cover the whole range of research from molecular pathogenesis to the development of new targeted therapies. The task of our group is to develop, validate, and apply an information platform that is tailored to the scale and multidisciplinary nature of the TRN. The integration of tissue, molecular, genetic, and clinical data into a common platform shall enable data sustainability and comprehensive analyses.

The aim of this paper is to introduce the special requirements that arise from a biomedical research network for the information platform and to discuss the resulting architecture blueprint. We want to share our experiences in using tools from the cancer Biomedical Informatics Grid (caBIG) initiative to build the system.

2. Methods

The 22 TRN projects are located at four major independent research institutions. Each project organizes its own research data. There is a considerable amount of data, distributed over the various institutions in different terminologies and in different formats. Standards, specifications, tools and standard operating procedures are necessary to ease the integration of biomedical research data on HCC. Our aim is to provide an efficient and secure environment to perform queries and analyses on integrated scientific information while respecting the distributed nature of the TRN.

2.1. The Pelican Architecture

The information platform pelican (**p**latform **e**nhancing **li**ver **ca**ncer **n**etworked research) is built in an iterative process. We started with a case study in two projects by analyzing the currently implemented way of data storage. In this study, we analyzed data structures, identified overlapping data and ambiguous data structures.

Further, we analyzed the applicability of open-source applications and specifications to our TRN. We found two major concepts of data storage for an integration platform: first, a central data warehouse into which data from all sources are loaded. I2b2 [7] is an example for a data warehouse used in biomedical context. The other concept is to federate data. That means, all data are kept separately but are made available for integrated analyses by using standardized interfaces. For example, caBIG [8] – the National Cancer Institute's (NCI) cancer Biomedical Informatics Grid – provides tools to build a federated system.

We decided to implement pelican as a service oriented architecture (SOA). As our base framework we chose components provided by the caBIG initiative. caBIG was established to improve cancer research by sharing, discovering, integrating, and processing disparate clinical and research data resources to improve cancer research. This includes the development of applications for data management and analysis, guidelines, and informatics standards. These tools are based on a grid architecture (caGrid) to link applications and resources in the caBIG environment [2].

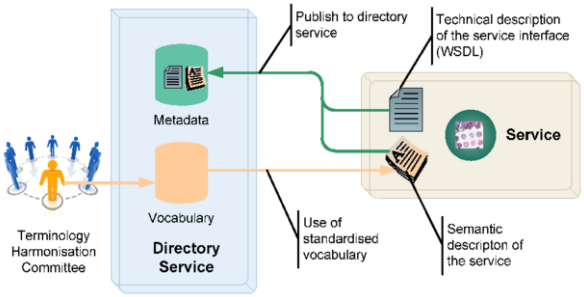

Figure 1. Semantics of pelican data services are described using a standardized vocabulary. Both technical metadata and semantic descriptions of the services are published to a directory service.

3. Results

In pelican, all data contributed by individual projects are transformed into data services using the Cancer Common Ontologic Representation Environment (caCORE) Software Development Kit (SDK). As shown in Figure 1, metadata are generated for all data services and published to a directory. The vocabulary used for the description has to be standardized throughout the TRN.

In addition to data services, analytical services are developed and made available within pelican. In the final version of pelican, researchers will be able to find data services hosting data necessary to answer their research question in the service directory. These data services are, together with analytical services, chained into a workflow that analyses the data of various sources and presents the results to the researcher. Figure 2 illustrates this process chaining concept. Further TRN projects can be easily added. Standard operating procedures and supporting tools will be developed to provide a smooth way of converting raw data as generated in the projects into data services conforming standards for pelican. In this process, it is especially important to apply the corresponding metadata correctly. Otherwise, it will be impossible to find the data sources in the directory and apply automated correlation algorithms.

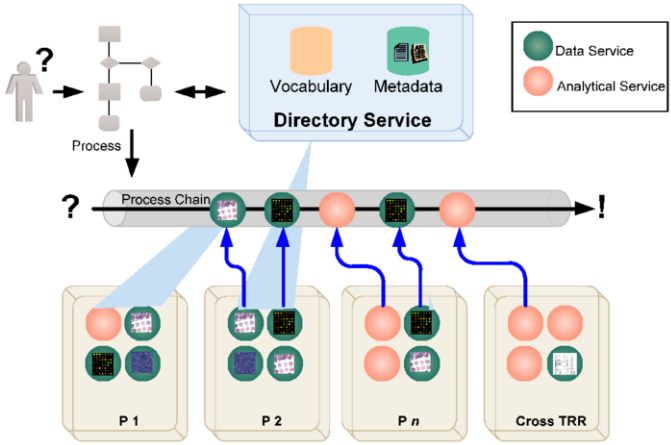

Figure 2. Individual services providing data or analytical services can be combined into a service chain. To do this, a researcher identifies the services of interest by using the service directory. The chain is executed by pelican and the results are returned.

3.1. TRN Specific Requirements

As a first step to assess the requirements for the pelican system, we conducted a survey among TRN project managers. For this purpose we developed a questionnaire consisting of 13 questions. For about one half of the questions standardized answers were provided with check boxes, the other half was free text. The survey covered various aspects of data usage, data confidentiality and intended use of the new system. The evaluation of the survey made obvious, that there is a strong concern among researcher about the confidentiality of data contributed to the system. These concerns were mostly about two aspects:

1. Researchers want to control who can access the data they contributed to the TRN. They want to keep the data confidential among specific project members or the whole TRN until they are published.
2. Getting access to the data of another project may lead to a significant advantage of somebody's own research. If this leads to a publication, rules have to be established and enforced how the contributors of the data are to be acknowledged e.g. by means of co-authorship.

3.2. Data Confidentiality and Intellectual Property

To address the TRN requirements, pelican architecture includes several confidentiality measures. Data services and as such data itself can be left under the control of the contributing project. Projects can define access control lists for their services if necessary. For the use of data generated by other projects, all TRN projects agreed on a set of rules. To support the enforcement of these rules, access to data is recorded by a comprehensive audit logging concept. Audit logs are monitored by the central project office on a regular basis. Our survey showed that 55% of the projects are only willing to share their data after data confidentiality concepts such as those proposed by us have been implemented in pelican.

3.3. Integration Platform

To test the architectural design of pelican, a prototype was built. It uses caCORE SDK to implement data services. However, it does not yet allow for dynamic process generation. Instead, a process for a specific research question is prepared statically. To answer this research question, it is necessary to correlate genomic microarray data of three data services provided by two projects. Data types used are array comparative genomic hybridization data (aCGH), methylation data and expression of coding and non-coding ribonucleic acid (RNA). In parallel, the service directory has been implemented and work on the standardized vocabulary has been started.

4. Discussion

When pondering whether the data warehouse or the federated approach would suit the needs of the TRN better, we chose to build a federated system. With this concept, it is possible for the individual projects to keep control over their data since data from different projects are encapsulated in distinct data services. Thereby, access control

mechanisms are easy to comprehend and to manage. In contrast, a data warehouse would usually combine all data in one database making it much harder to apply access permissions on an individual basis and communicate these settings to the projects.

Using the pelican prototype, we were able to demonstrate, that it is possible to integrate data of different TRN research projects using a SOA based on caBIG components. We were able to statically correlate several genetic data sources and thus support the researchers of two projects. Further, we started to implement the metadata directory and the implementation of caGRID [2].

The security concept designed for pelican is accepted throughout the TRN, as our survey substantiates. However, further work needs to be done regarding the user interface to ensure user acceptance: pelican should be as easy to use as the tools currently used by the researchers.

To enable the dynamic composition of process chains, a workflow engine has to be added to pelican. Candidates for this are Business Process Execution Language (BPEL) based engines or the Taverna workflow management system [9]. Finally, we need to examine other caGRID enabled tools provided by the caBIG initiative to find out, if those can complement pelican to further improve the TRN research.

Acknowledgements: The authors would like to thank the German Research Foundation (DFG) for funding SFB/TRR 77 – "Liver Cancer. From molecular pathogenesis to targeted therapies."

References

[1] Knaup P, Ammenwerth E, Brandner R, et al. Towards clinical bioinformatics: advancing genomic medicine with informatics methods and tools. Methods Inf Med 2004; 43(3):302–7.
[2] Oster S, Langella S, Hastings S, et al. caGrid 1.0: an enterprise Grid infrastructure for biomedical research. J Am Med Inform Assoc; 15(2):138–49.
[3] Madhavan S, Zenklusen J, Kotliarov Y, Sahni H, Fine HA, Buetow K. Rembrandt: helping personalized medicine become a reality through integrative translational research. Mol. Cancer Res 2009; 7(2):157–67.
[4] Lee TJ, Pouliot Y, Wagner V, et al. BioWarehouse: a bioinformatics database warehouse toolkit. BMC Bioinformatics 2006; 7:170.
[5] Hart RK, Mukhyala K. Unison: an integrated platform for computational biology discovery. Pac Symp Biocomput 2009:403–14.
[6] Schork NJ. Genetics of complex disease: approaches, problems, and solutions. Am. J. Respir. Crit. Care Med 1997; 156(4 Pt 2):S103-9.
[7] i2b2: Informatics for Integrating Biology & the Bedside [cited 2011 Apr 19]. Available from: URL:https://www.i2b2.org/.
[8] Welcome to the caBIG® Community Website — [cited 2011 Apr 19]. Available from: URL:https://cabig.nci.nih.gov/.
[9] Tan W, Missier P, Foster I, Madduri R, Goble C. A Comparison of Using Taverna and BPEL in Building Scientific Workflows: the case of caGrid. Concurr Comput 2010; 22(9):1098–117.