@neurIST – Towards a System Architecture for Advanced Disease Management through Integration of Heterogeneous Data, Computing, and Complex Processing Services

H. Rajasekaran[1], L. Lo Iacono[1], P. Hasselmeyer[1], J. Fingberg[1], P. Summers[2], S. Benkner[3], G. Engelbrecht[3], A. Arbona[4], A. Chiarini[5], C.M. Friedrich[6], M. Hofmann-Apitius[6], K. Kumpf[6], B. Moore[7], P. Bijlenga[8], J. Iavindrasana[8], H. Mueller[8], R.D. Hose[9], R. Dunlop[10], A.F. Frangi[11]

[1] IT Research Division, NEC Laboratories Europe, NEC Europe Ltd., Sankt Augustin, Germany
rajasekaran@it.neclab.eu
[2]University of Oxford, Oxford, UK
[3]Institute of Scientific Computing, University of Vienna, Vienna, Austria
[4]GridSystems S.A., Palma de Mallorca, Spain
[5]B3C BioComputing Competence Centre, Bologna, Italy
[6]Fraunhofer Institute SCAI, Sankt Augustin, Germany
[7]IDAC Ireland Limited, Ireland
[8]Geneva University Hospitals, Geneva, Switzerland
[9] University of Sheffield, Sheffield, UK
[10] Infermed, UK
[11]Information & Communications Technology Department, Universitat Pompeu Fabra, Barcelona, Spain

Abstract

This paper presents the system architecture of the @neurIST project, which aims at supporting the research and treatment of cerebral aneurysms by bringing together heterogeneous data, computing and complex processing services. The architecture is generic enough to adapt it to the treatment of other diseases beyond cerebral aneurysms. The paper describes the generic requirements of the system and presents the architecture, applications and middleware technologies used to realise the system and highlights the innovations in @neurIST.

1. Introduction

Like in other knowledge intensive fields, the volume of data describing human disease processes, including their understanding, diagnosis, and management, is growing exponentially. The data are increasingly heterogeneous, including text, images, and other symbolic structures. Data are also diverse in context, from global guidelines based on broad epidemiological studies, through knowledge gained from disease-specific scientific studies, both in vitro and in vivo, to individual patient-specific data.

The data spans scales from molecular, through cellular, to tissue, organ and patient representations, and are increasingly inter-related in part due to recent breakthroughs in our understanding of disease processes through functional genomics studies.

The huge volume of this information and its rate of growth represent an unprecedented data management challenge. In particular it is often impossible for an individual, whether a clinician responsible for patient management, or a physicist or engineer developing a new generation of imaging or interventional devices, to understand and assimilate this knowledge. It is increasingly clear that new methods are required to manage, integrate and interrogate the data in a manner that is accessible to the end user.

This paper presents the architecture of a system developed through the @neurIST European Integrated Project to tackle this problem. The paper highlights the challenges that need to be overcome and the innovative elements that have been designed to contribute to the integration of heterogeneous data and computing resources, and to enable and support a number of integrative application suites that aim at underpinning comprehensive risk assessment and treatment planning procedures. The system also supports tools for basic researchers in their effort to link genetic to phenotypic information.

2. The @neurIST project

@neurIST is an Information Society Technologies (IST) Integrated Project funded within the European Commission's (EC) Sixth Framework Programme. The @neurIST system focuses on supporting the research and treatment of cerebral aneurysms. The project aims at building a distributed IT infrastructure that consolidates complex data from multiple sources, and enables personalized patient management (i.e. data capture, referral, decision support, treatment planning), as well as clinical research in cerebral aneurysms [1,2]. The consortium brings together 30 multi sectorial partners representing hospitals, universities, research institutes and the industry across Europe.

At the heart of the project is the clinical problem of managing unruptured cerebral aneurysms and associated research into risk factors. Currently, aneurysm diagnosis and treatment relies on the interpretation of numerous pieces of information, which are recognized to give an incomplete picture of the disease. In part, this information comes from the patients themselves in the form of radiographic images, family history and physiological measurement, and in part it is derived from the understanding of the available medical literature and the past experience of the patients' clinicians. One shortcoming is that the process is subjective and at the same time the scales of risk involved are difficult to communicate. This places the clinicians and often the patient in the situation of taking decisions with significant risk involved with an appreciation of the situation that is rarely optimal. Moreover, the data is inherently incomplete and, to cover all contingencies, patients may be subjected to tests and undergo treatments which are of minimal real benefit or opt against potentially life-saving treatment simply due to a lack of knowledge, information, or understanding.

The same issues apply to many other diseases. The end result is that the quality of care is sub-optimal, and its cost is increased. While there is undoubtedly a cost in human terms, there is also a quantifiable financial cost. This cost is recognized as substantial relative to other disorders of similar incidence and risk by all major healthcare players. This cost can form the basis for a long-term business model. @neurIST aims at keeping this cost low and raising the quality of diagnostics and treatment by building a distributed IT infrastructure, following a Service Oriented Architecture (SOA) model, which integrates Grid service technologies to bring together the different stakeholders involved in the research and treatment of a particular disease.

2.1. @neurIST Scientific and Functional Requirements

Though the principle focus of @neurIST is cerebral aneurysms, the implications of the @neurIST solution are far wider in scope. The point-of-care integration of imaging, genomic, and clinical results with epidemiological and other data is applicable in all fields of medicine. The generic processes that occur repeatedly in most studies and treatment processes of diseases are

- *Obtaining relevant clinical information about each patient's clinical problem/s.* Clinical data is often widely distributed, particularly if a patient has undergone a health care journey involving diagnostic and treatment procedures in multiple medical institutions. The clinician may require additional data at successive stages in the patients' treatment process making the data itself a dynamic entity.
- *Providing clinical decision support, which provides guidance on best practice.* Informed decisions about further testing, treatment options and procedures are required at every step. Simply consolidating the numerous pieces of information, and recording their interpretation is one aspect of this. Providing context so that appropriate best practice and recent findings are adequately considered and can be reviewed by all involved in the clinical management is a second.
- *Offering simulation services*, which allow clinicians and patients to assess complex treatment options. These services can allow for more efficient procedures by allowing rehearsal and ensuring the treatment device of choice is in stock for the procedure. These services can also enable new treatment solutions to be devised and tested by medical device manufacturers.
- *Creating normalized population-based datasets*, where clinical data is gathered during the process of care, resulting in the reduction of the cost of population studies and an increase in their scope. At the same time, the information can provide new insights into management, as well as an audit mechanism for the quality of care.
- *Providing knowledge discovery services*, by facilitating the acquisition and processing of complex data sources, such as the vast array of medical literature or publicly accessible genomics databases.

Implementing these requirements involves addressing issues such as

- *Combining heterogeneous data sources*, which are distributed geographically, encoded in different formats and may involve different languages and nomenclatures.
- *Providing sufficient computing power* for complex simulations and other data processing tasks.
- *Maintaining the privacy of patient data and securing the entire IT infrastructure* to ensure compliance with legal regulations and to ward off malicious attacks.

3. The @neurIST System Architecture

The @neurIST system has essentially two modes or cycles of operation. In the first mode, it operates as an integrative decision support system to help in diagnosis, risk assessment and treatment planning integrating complex and multifactorial information. In a second mode of operation, it works as a system enabling *in silico* research and understanding of the underlying causes of disease and to link the genetic and phenotypic evidence available in large-scale text databases so that new knowledge can be extracted, structured and transposed for its exploitation in the decision support operation cycle. While the decision support system largely deals with actual patient data, the research system has access only to anonymized data sets. These two cycles are underpinned by several integrative applications, which are organised into suites that help in data collection, simulation, risk assessment, etc., and system components that connect databases and services distributed at different sites. The system components take care of the process involving security, data access, data transport and computing.

The system is developed using multiple technologies and deployed across a wide geographic distribution. @neurIST adopts the SOA approach to integrate such a diverse system and uses open standards and technologies from the internet, Grid and medical worlds. Figure 1 shows the layered view of the reference architecture of @neurIST with their constituent components. The reference architecture is explained by focusing its application to the treatment of aneurysms. The @neurIST system can be divided into three layers – application, middleware and resource.

3.1. Application layer

The application layer consists of the applications that provide end users with functionalities for research and clinical treatment of aneurysms.

The **@neuLink** application suite is targeted at basic research users within @neurIST, essentially geneticists and epidemiologists. Its objective is to link genetic information with disease information. The information is gathered from public bioinformatics databases and from the Biomedical Info Structure (BioIS is the patient database system in the clinical centres in @neurIST; see section 3.3). In this way it supports the clinical decision process with evidence from analyzed experiments and knowledge discovery from other data sources. The aggregated information is provided via a web-application to a human user, as well as via the @neurIST data access infrastructure, to support other application suites. The main modules support the search for candidate genes and associated genetic variations (Single Nucleotide Polymorphisms (SNP)) from text and from the aggregated protein-protein interaction database PIANA [3]. The suite supports the interpretation of Microarray experiments and provides results from applying data mining methods for SNP association studies and risk assessments from the anonymized patient population. Initial results on knowledge discovery were presented in [4].

The **@neuFuse** suite receives datasets from medical imaging and biomedical instrumentation, as well as derived data such as indicator fields obtained from simulation services. The application enables medical

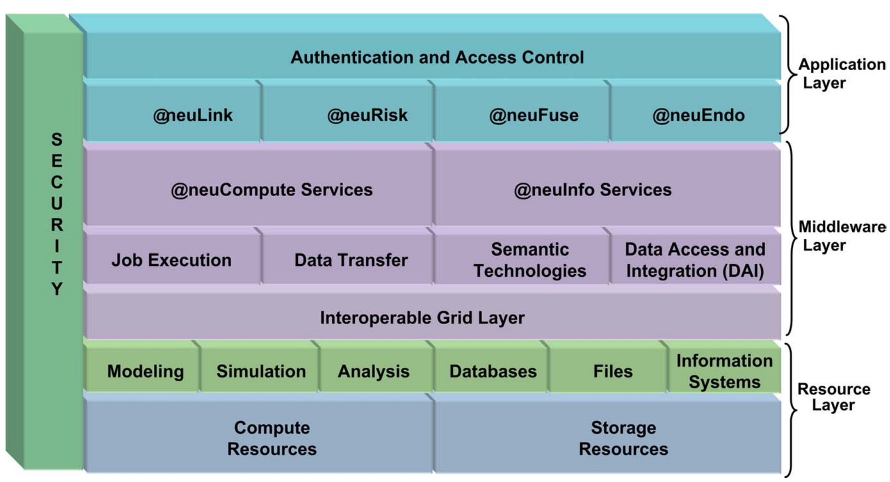

Figure 1 @neurIST Reference Architecture - layered view derived datasets to be organized in time and space to be interactively visualized independently or simultaneously and to be edited. The editing process changes existing datasets or creates new datasets depending on the application logic implemented for each data type. This mechanism protects the integrity of the original diagnostic data, an aspect that is vital in clinical scenarios. @neuFuse provides information to be used by clinical users as well as researchers via the @neuInfo middleware. This information is also accessed by other application suites.

@neuEndo is a tool that uses cutting edge Finite Element Analysis (FEA) and Computational Fluid Dynamics (CFD) technology to assess mechanical and flow performance of stents used for treatment of aneurysms. It is aimed at both industrial and clinical end-users. Industrial users use it as a design tool during stent development and clinicians use it to assess the suitability of commercially available stents for particular patient conditions. @neuEndo currently has two distinct applications. One looks at the stent from a structural perspective and one looks at the aneurysm from a flow perspective. Ultimately, the plan is to combine the two disciplines and look at the interaction between stent, vessel and blood.

@neuRisk forms the core of the integrative decision support system. It integrates all available information for identifying patients who are at high risk for rupturing asymptomatic cerebral aneurysms. It makes use of the functions or knowledge generated in other application suites to include genetic and simulation information in the risk assessment process. A key innovative element in @neuRisk is its ability to integrate information from a wide range of origins through an underlying formalism called argumentation which allows to reason under uncertainty and provide supporting and non-supporting evidence behind alternative treatment options. Therefore, in addition to being useful for clinicians, this system is particularly useful for counselling patients and facilitating their trading off benefits and risks in deciding for their own treatment path. @neuRisk will eventually improve the decision-making processes and reduce current over-treatment.

3.2. Middleware Layer

The middleware layer in @neurIST mediates between the application layer and the resource layer, providing data access, data staging, semantic mediation, and grid computing services. Through this mediation the applications do not need to be aware of the exact physical or logical sources of the data or computing resources. The middleware layer comprises the @neuInfo and @neuCompute infrastructures, offering data and compute services to the application suites. Data and compute services are based on standard Web Services technologies. They are defined via the Web Services Description Language (WSDL) and are securely accessed through the exchange of SOAP messages.

@neuCompute enables service providers to virtualize High Performance Computing applications available on clusters or other parallel hardware as compute services that can be accessed on-demand by applications. Compute services hide the details of their execution environment, providing abstract interfaces for managing the job execution. The core of @neuCompute consists mainly of two software developments, Fura [5] and Grid-Enabled Medical Simulation Services **(**GEMSS) [6]. @neuCompute comprises two layers, an Intra-Grid layer which focuses on high-throughput computing and cycle harvesting based on Fura, and an Extra-Grid layer, which offers generic WSDL interfaces to clients as developed within GEMSS. This outer GEMSS layer was significantly re-factored for @neurIST adopting a more flexible component-based approach.

Internally, compute services are composed of basic service components including a data transfer component handling the transfer of input and output data between the client and the service, an application execution service component offering operations for executing application jobs on the associated compute resource(s) (possibly relying on Fura) and for querying the status of an application job, and a data staging service component if input and/or output data should be transferred directly between services rather than between a client and a service. The Quality of Service (QoS) component offers operations to support the dynamic negotiation of QoS guarantees. Finally, an error recovery service component can be used in order to provide support for error recovery based on a checkpoint/restart mechanism.

@neuInfo offers a generic framework that supports the provision and deployment of data services. Data services virtualize heterogeneous scientific databases and information sources as Web services, enabling transparent access to and integration of relational databases, XML databases and flat files.

Data services hide the details of accessing distributed data sources, resolving heterogeneities with respect to access language, data model and schema. Internally, they utilize OGSA-DAI [7] (Open Grid Services Architecture - Data Access and Integration) for accessing actual data sources, while providing the same access patterns, client bindings, transfer protocols and security mechanisms as compute services. Data services can be set up in different configurations all providing the same interface to clients. While basic data services provide access to a single data source, data mediation services support transparent access to multiple data sources by establishing a global virtual schema.

Data mediation, which is initially based on handwritten mapping schemas, is now being enhanced based on semantic technologies in order to reduce the integration effort. Using the @neurIST ontology [8], individual data sources will be annotated semantically. From the semantic annotation of individual data schemas (tables, columns), mapping schemas required for query reformulation of data mediation services will be generated automatically. Moreover, data mediation services will be enhanced in order to support queries at the level of ontology concepts [9].

3.3. Resource Layer

The resource layer consists of the computational resources and databases. The compute resources are offered by service providers to run compute intensive processes in the different application suites of @neurIST. The databases offer storage resources to store simulation results and patient data. In particular, the databases in the clinical centres that store patient information form a part of the BioIS in @neurIST. It connects the @neurIST clinical centres within the @neurIST system. It is an @neurIST specific implementation, which is driven by the need to gather clinical data from multiple clinical centres for research and treatment purposes [10].

The BioIS exchanges information via the @neuInfo system (distributed queries and updates). This includes requests for clinical data, as well as patient-specific data outputs from other @neurIST services. Clinical information will remain at its source within the clinical information systems (CIS) of participating centres. The BioIS sub-system interfaces with the CIS databases. The BioIS supports two architectural styles: 1) the anonymized database model (ANO), which is based on a dedicated database containing a pseudonymised copy of the patient records. 2) An on-the-fly (OTF) model that communicates directly with the CIS. Figure 2 shows the two models employed by the BioIS.

3.4. Security in @neurIST

Security is an essential requirement in any functional system and in the case of @neurIST it is imperative, due to the sensitive nature of data that is handled by the system involving multiple stakeholders. Any compromise to the patient data in terms of breach of privacy, disclosure to unauthorized personnel, or misplaced data will have a serious impact on the credibility and effectiveness of the system to treat patients.

The integrated and comprehensive @neurIST security framework includes various functional components to ensure that patient data is made available only to authorized personnel, preserves the privacy of the patient at all times and provides an efficient infrastructure to authenticate and authorize @neurIST personnel and applications across multiple domains in accordance with legal and ethical requirements.

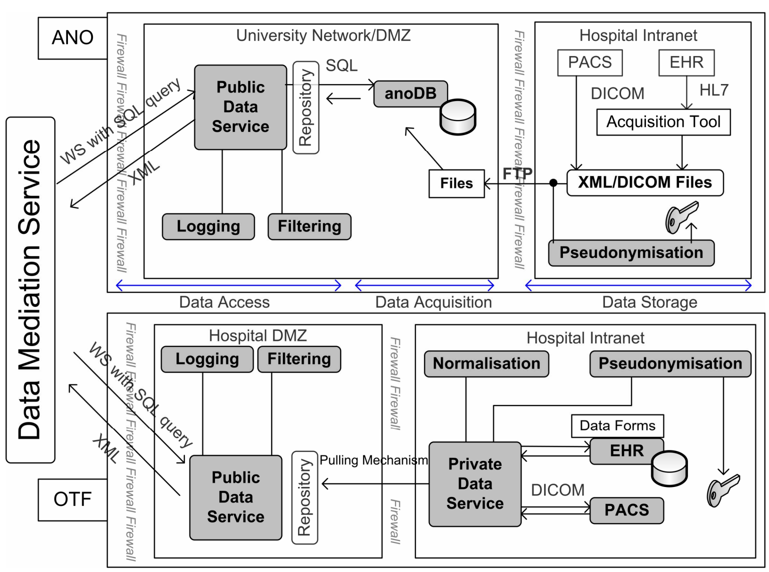

Figure 2: ANO and OTF architectural solutions

The pseudonymisation system depersonalizes the patient data required for the study including textual data and medical images while maintaining the possibility to re-contact patients in the event of relevant results being obtained in the course of the study.

The @neurIST trust and VO (Virtual Organization) model gathers all the participating entities such as clinics, research institutes and service providers into one large VO, whereas the management of the VO is realized in a decentralized manner allowing the efficient setup of purpose-specific federations. A credential issued to @neurIST users should be recognized (either directly or through the construction of adequate trust paths) by all participating entities in the system to grant access to services and data in the system. Though the users' credentials are issued and administered by the local entity to which they belong, the access to services and data is controlled based on policies set by the entities that own them. Further components constituting the @neurIST security framework include filtering and delegation mechanisms for access control as well as logging and monitoring functionalities for usage control, intrusion detection and post-incidents analysis.

4. Discussion and Conclusions

In this paper we present the @neurIST architecture that supports applications and services to help in the research and clinical treatment of aneurysms. The architecture is designed to be independent of the disease being treated such that most of the architecture can be reused and new developments need to be done only on the disease specific aspects of a) the knowledge structures (CRIM, ontology and risk model formulation) and b) the disease relevant aspects of the application suites.

Among the innovations brought about in by the project are incorporating research into the clinical decision support system, including offering complex FEA and CFD simulations. These introduce the scope to try out possible treatment options and devise the appropriate treatment plans. The project also offers real world models to industrial partners to improve their stent designs and a semantically mediated, anonymized dataset derived from the clinical data of the participating clinics for the @neurIST researchers with the appropriate security measures for their research. The next steps in the project are further integration of the individual components and running clinical trials to evaluate the design and performance of the @neurIST system.

5. Acknowledgment

This work was generated in the framework of the @neurIST Integrated Project, which is co-financed by the EC through contract no. IST-027703.

6. References

[1] A. Arbona, S. Benkner, G. Engelbrecht, J. Fingberg, M. Hofmann, K. Kumpf, G. Lonsdale, A. Woehrer. A service-oriented grid infrastructure for biomedical data and compute services. *IEEE Trans Nanobioscience*. 2007 Jun;6(2):136-41.

[2] A. Arbona, S. Benkner, J. Fingberg, AF. Frangi, M. Hofmann, DR. Hose, G. Lonsdale, D. Ruefenacht, M. Viceconti, Outlook for grid service technologies within the @neurIST eHealth environment. *Stud Health Technol Inform*. 2006;120:401-4.

[3] R. Aragues, D. Jaeggi, B. Oliva, PIANA: protein interactions and network analysis. *Bioinformatics*, 15;22(8):1015-7

[4] C.M. Friedrich, M. Hofmann-Apitius, R. Dunlop, I. Chronakis, M.C.J.M. Sturkenboom, R. Risselada, B. Oliva and F. Sanz, Initial Results on Knowledge Discovery and Decision Support for Intracranial Aneurysms, *HEALTHINF 2008, INSTICC*, 2008, 265-272.

[5] Fura; http://fura.sourceforge.net/

[6] S. Benkner, G. Berti, G. Engelbrecht, J. Fingberg, G. Kohring, SE. Middleton, R. Schmidt.; GEMSS: grid-infrastructure for medical service provision. *Methods Inf Med*. 2005;44(2):177-81. http://www.it.neclab.eu/gemss/

[7] OGSA-DAI; http://www.ogsadai.org.uk/

[8] M. Boeker, H. Stenzhorn, K. Kumpf, P. Bijlenga, S. Schulz, S. Hanser, The @neurIST Ontology of Intracranial Aneurysms: Providing Terminological Services for an Integrated IT Infrastructure. *AMIA 2007 Annual Symposium (AMIA 2007)*, Chicago, USA, November 2007.

[9] K. Kumpf, A. Woehrer, S. Benker, G. Engelbrecht, and J. Fingberg, A Semantic mediation architecture for a clinical Data Grid. Grids for Bioinformatics and Computational Biology, *Wiley & Sons series on bioinformatics* (2007), in press.

[10] J. Iavindrasana, A. Depeursinge, P. Ruch, S. Spahni, A. Geissbuhler, H. Müller, Design of a Decentralized Reusable Research Database Architecture to support data acquisition in large research projects. *Medinfo. 2007*;12(Pt 1):325-9.