Towards Collaborative Data Management in the VPH-Share Project

Siegfried Benkner[1], Jesus Bisbal[2], Gerhard Engelbrecht[2], Rod D. Hose[3], Yuriy Kaniovskyi[1], Martin Koehler[1], Carlos Pedrinaci[4], and Steven Wood[5]

[1] Faculty of Computer Science, University of Vienna, Austria
[2] Center for Computational Imaging & Simulation Technologies in Biomedicine, Universitat Pompeu Fabra, Barcelona, Spain
[3] Department of Cardiovascular Science, Medical Physics Group, University of Sheffield
[4] Knowledge Media Institute, The Open University, Milton Keynes, UK
[5] Dept. Medical Physics, Royal Hallamshire Hospital, Sheffield, UK

Abstract. The goal of the Virtual Physiological Human Initiative is to provide a systematic framework for understanding physiological processes in the human body in terms of anatomical structure and biophysical mechanisms across multiple length and time scales. In the long term it will transform the delivery of European healthcare into a more personalised, predictive, and integrative process, with significant impact on healthcare and on disease prevention. This paper outlines how the recently funded project VPH-Share contributes to this vision. The project is motivated by the needs of the whole VPH community to harness ICT technology to improve health services for the individual. VPH-Share will provide the organisational fabric (the infostructure), realised as a series of services, offered in an integrated framework, to expose and to manage data, information and tools, to enable the composition and operation of new VPH workflows and to facilitate collaborations between the members of the VPH community.

Keywords: virtual physiological human, healthcare infrastructure.

1 Introduction

The Virtual Physiological Human Initiative (VPH-I) from the European Commission aims to provide a systematic framework for understanding physiological processes in the human body in terms of anatomical structure and biophysical mechanisms at multiple length and time scales. Multiple projects are funded and try to meet specific objectives addressing data integration and knowledge extraction systems or patient specific computational modeling and simulation. To achieve these objectives a combined data/compute infrastructure will need to be developed.

M. Alexander et al. (Eds.): Euro-Par 2011 Workshops, Part I, LNCS 7155, pp. 54–63, 2012.

The VPH-Share project[1] has been funded within the VPH initiative and will provide a systematic framework for the understanding of physiological processes in the human body. A long term objective is to transform the European healthcare into a more personalised, predictive, and integrative process with significant impact on healthcare and on disease prevention. The project will provide an integrated framework, realised as set of services, to expose and manage data, information and tools, to enable the composition and operation of new VPH workflows and to facilitate collaborations between the members of the VPH community. The project consortium comprises 21 partners from the European Union and New Zealand including data providers, providing data sources from individual patients (medical images and biomedical signals), research institutes, universities, and industry.

The project addresses four flagship workflows from European projects which provide existing data, tools, and models driving the development of the infostructure and pilot the applications. The flagship workflow include a workflow from the @neurIST project[2], dealing with the management of unruptered cerebral aneurysms and associated research into risk factors. The euHeart[3] workflow supports integrated cardiac care using patient-specific cardiovascular modeling and the VPHOP workflow[4] is in the domain of osteoporotic research. The fourth flagship workflow, Virolab[5], drives a virtual laboratory for decision support for the treatment of viral diseases. By covering these workflows the VPH-Share project aims at the provisioning of a generic data management and computational infrastructure for supporting generic VPH workflows.

The main focus of the project is the provisioning of a patient avatar which can be defined as a coherent digital representation of a patient. The provisioning of a patient avatar will rely on the DIKW hierarchy[6] promoted by the ARGOS Observatory [1]. On the lowest layer, the DIKW pyramid, proposes data including instantiations of measurements. By utilizing data as input to diagnosis it becomes information which can be cognitively processed by means of knowledge. If knowledge becomes confirmed and accepted, it is called wisdom. By following these paradigm new data sources need to be established and made widely available through an appropriate infrastructure, including a data management platform, which we call data infostructure.

In the following, we clarify some terminology used in the rest of this paper. By the term 'infrastructure', we mean the raw data, and the tools and services that operate on them (for example to access, to transfer, to store) without any

[1] VPH-Share: `https://www.biomedtown.org/biomed_town/vphshare/reception/website/`

[2] @neurIST: Integrated biomedical informatics for the management of cerebral aneurysms, `http://cilab2.upf.edu/aneurist1`

[3] EuHeart: Integrated cardiac care using patient-specific cardiovascular modeling, `http://www.euheart.eu`

[4] VPHOP: the Osteoporotic Virtual Physiological Human, `http://www.vphop.eu`

[5] ViroLab: a virtual laboratory for decision support in viral diseases treatment, `http://www.virolab.org`

[6] DIKW hierarchy: `http://en.wikipedia.org/wiki/DIKW`

understanding of the content of the data, and the hardware resources that are used in all data and modelling operations. We use 'infostructure' to describe the systems and services that VPH-Share will develop to transform data into information and thence into knowledge.

2 VPH-Share Infostructure

The DIKW hierarchy described in the previous section inspired the vision for the infostructure the VPH-Share project aims to build. The main components of this infostructure are presented in Figure 1 as a layered architecture. It illustrates the generation of new (medical) wisdom and the respective tools and services to be developed/used for this - from data to knowledge, through the VPH-Share enabled infrastructure. New, validated VPH models will thus be developed and integrated into so-called 'patient-centred computational workflows' (detailed in Section 2.2).

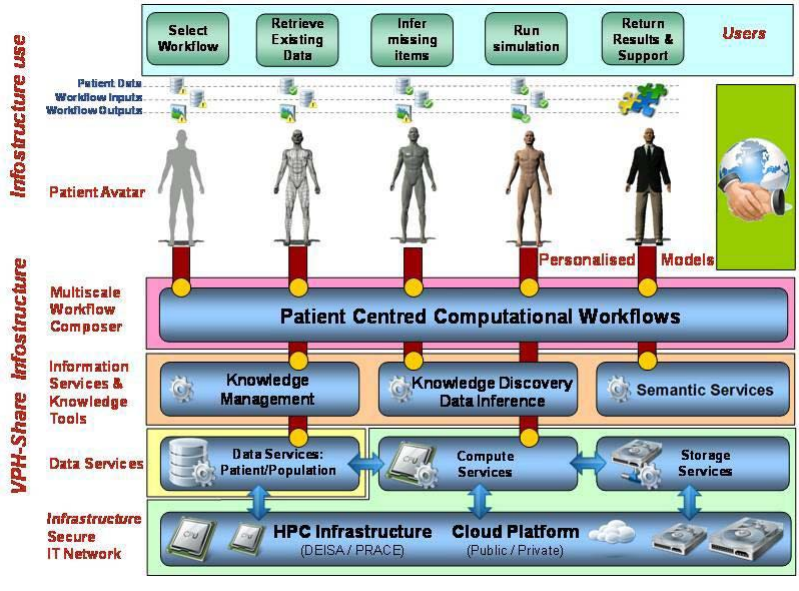

Fig. 1. VPH-Share Architecture

More specifically, starting at the bottom of Figure 1, the architecture includes the lowest level of services that provide access to computational infrastructure and manage execution of VPH-Share operations. It is foreseen that both, the Cloud computing paradigm as well as high performance computing services, will be available in order to address the wide diversity of challenges in the VPH. From the data management perspective, this first layer is mainly aimed at the management of large files. More structured (e.g. relational, xml) data is managed by a set of specific data services to contribute, access, distribute, and annotate these type of data sets.

On top of these services, advanced semantic services are added to facilitate the sharing and re-use of these datasets. In addition, data inference strategies and engines are also added, in order to exploit the wealth of knowledge hidden within the vast amounts of data that will be stored within this infostructure. In that respect, it must be noted that biomedical datasets are inevitably incomplete, thus these services will generate (or, rather, 'infer') this input from other relevant data which is available.

The architecture also provides a unified, but modular, user interface to all available services.

Concrete realisation of the vision is effected in four patient-centred computational workflows, as defined in Section 2.2. This ensures that all advances, tools and services developed within the project are at all times fit for purpose and meaningful to the biomedical researcher.

2.1 Patient Avatar

The VPH-Share project has introduced a central concept, referred to as the *Patient Avatar*. This is an evolving concept which has received different names since its initial conception. For example, it was called the Virtual Patient Metaphor, in the @neurIST project [3], and more recently in the Network of Excellence (NoE) is called the *Digital Me* [2]. Following the strategic vision defined by the NoE, the patient avatar is described as: "a coherent digital representation of each patient that is used as an integrative framework for the consolidation within the European research system of fundamental and translational Integrative Biomedical Research and the provision to European Citizens of an affordable Personalised, Predictive, and Integrative Medicine".

For a concrete realisation of the patient avatar the project provides the means to be specific about which information it must contain to be relevant to specific contexts, a concept that will be explicitly and directly tested within each of our VPH-Share workflows. At the least personalised level this avatar will contain population averages, or even best guesses, for all information items. The progression from the silhouette to the clothed man in Figure 1 illustrates the personalisation of the avatar as the VPH-Share data inference services operate on the information that is available about the individual to refine the estimates of those data items (and their likely ranges) that have not been measured or recorded.

2.2 VPH-Share Workflows

The concepts associated with the construction of an infostructure can become very abstract, and the project recognised the danger of trying to impose on the community a solution that might be conceptually elegant, computationally efficient, and even robust, but which may ultimately be very difficult to use by typical VPH researchers. To address this issue we have selected four driving patient-centred computational workflows, which we would suggest represent some of the best from completed or running ICT projects in the 6th and 7th

Framework Programmes, namely @neurIST, euHeart, VPHOP and Virolab, to serve as the empirical basis and benchmarks for the support structure to be developed by this proposal. It can be claimed that together they encapsulate the breadth of challenges presented to the VPH researcher.

Our use of the flagship workflows to guide the development of the infostructure, and to pilot its application, is consistent with the VPH NoE Vision document's recommendation that "all progress in the VPH must be driven and motivated through associated complex clinical workflows" [2]. In spite of the variety of problems the VPH as a whole addresses, there is a relative small number of possible workflows that are being developed to address the general problem of producing personalised, quantitative, and predictive models. This observation creates an opportunity for standardisation of methods and tools, which must constitute the backbone of this infostructure. Its construction is the ultimate goal of the VPH-Share project.

2.3 VPH-Share Cloud Infrastructure

The VPH-Share project will provide a Cloud infrastructure facilitating access to data and compute resources needed for data hosting and the execution of applications. On top of the Cloud infrastructure, VPH-Share services including data, semantic, and compute services, as well as workflows, will be hosted on demand. On the data side, a key requirement is to enable data hosting locally at the data provider's site, as this is the key requirement of many clinical institutions. Since some compute services and workflows have demanding compute requirements, access to HPC e-infrastructures will be integrated too.

On top of these requirements, there is a need for public and private Cloud environments, as well as access to HPC resources. The main goal is not to implement low-level Cloud middleware services, but rather to built a flexible Software as a Service (SaaS) environment on top of existing Infrastructure as a Service (IaaS) solutions. The infrastructure will provide easy deployment and execution of scientific applications and on-demand Cloud resource management. The Cloud infrastructure will include a policy driven security framework which ensures that the information exchange between VPH users and the services and data stored in the Cloud is secure and reliable.

3 VPH-Share Data Infrastructure

The VPH-Share data infrastructure aims at creating a unified data management platform supporting the efficient management and sharing of biomedical information consented for research. The platform comprises generic services and protocols to enable data holders to manage, provide, and share the information. The design of the data management platform follows an incremental Extract-Transform-Load (ETL) process that allows the provisioning of an evolving platform that can be extended as new information sources become available. Using semantic data integration technologies, the platform supports on-demand customised views on the available information [4].

The data infrastructure provides different types of services supporting access and integration of federated data sources. The focus of the infrastructure is the provisioning of relational data sources that are available in the form of relational databases or as files following a relational schema (e.g. CSV files). The services support querying of the data sources via relational as well as semantic concepts and provide a consistent interface to the other software components involved in the project. The services are exposed on top of the VPH-Share Cloud-based resource infrastructure.

The data infrastructure provides a uniform data management platform on top of services achieving the following objectives:

3.1 VPH-Share Data Sources

The VPH-Share project identified multiple VPH-relevant data sources which will be supported and Cloud-enabled by the data infrastructure. These data sources include clinical, research and simulation data sources, accessible via the data management platform. The VPH-Share project will discuss the requirements for different data sources, design patterns, as well as data schemes together with the VPH Network of Excellence (NoE). To integrate these and new data sources into the data management platform, there is a need for on-demand data transformation.

Data exposed within VPH-Share will be employed in the context of the infostructure and will be exposed to VPH-Share stakeholders following security and privacy requirements. Datasets that have been identified for the provisioning via the data infrastructure include data sources from the European projects @neurIST and ViroLab, as well as the NHS IC database [7], and the STH Cardiac data set.

The @neurIST dataset holds information in the domain of cerebral aneurysm research, including images, comprehensive demographic, and physiological information obtained from six European member states. The ViroLab data set includes several thousand records associated with HIV/AIDS research including genomic sequences, genotypes, treatment history, clinical and demographic data. The dataset from the NHS IC contains a range of national health and social care datasets that describe the demographics, lifestyles, burden on the health and social care system and interaction with this system. A longitudinal data set including cardiac data is utilized at the Sheffield Teaching Hospital (STH) and can possibly be utilized during the project as well. A number of additional data sets have been identified and it has yet to be decided if they can be included in the data management platform based on the data holders requirements and legal restrictions.

3.2 Data Services

The VPH-Share project will provide a generic data management and integration framework that supports the provisioning and deployment of data services.

[7] NHS Information Centre, `http://www.ic.nhs.uk/`

The data service infrastructure will enable the virtualization of heterogeneous scientific databases and information sources as Web services which do allow transparent access to and integration of relational databases, XML databases and flat files. The development of data services is based on the Vienna Cloud Environment (VCE) [5],[6] and the @neurIST data service infrastructure [7] and utilizes advanced data mediation and distributed query processing techniques. Data services hide the details of distributed data sources, resolving heterogeneities with respect to access language, data model and schema. These services comprise data access services to expose data sources via a Web Service interface. Additionally, data mediation services are provided in order to transparently combine different data sources and data access services in a mediated fashion as high-level services. Data mediation services preserve the autonomy of underlying data sources and ensure always up-to-date data, both key requirements of the project. A customised set of these services forms the basis for an on-demand dataspace which can be utilized by the workflows and the end users.

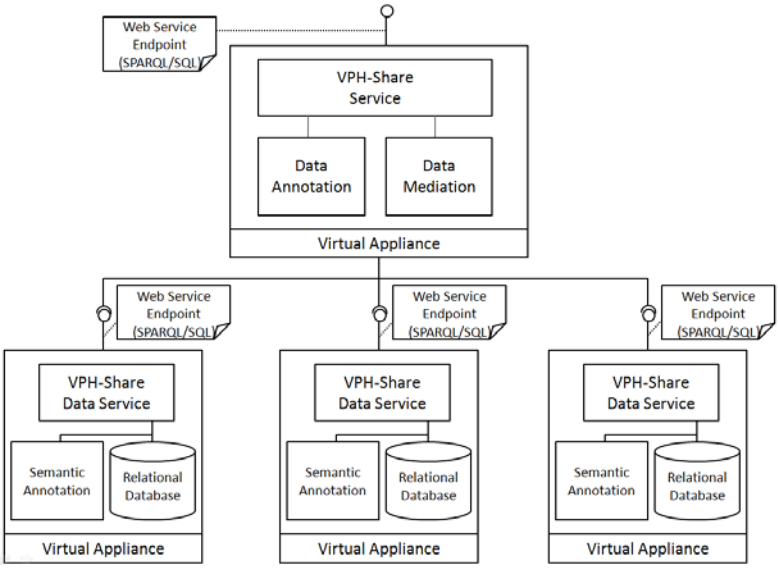

Fig. 2. VPH-Share Data Services: Data Services are hosted in virtual appliances and expose data sources as Web service endpoints

The data service infrastructure, as outlined in Fig. 2, is being built on top of state-of-the art Web service technologies. Data access services provide a uniform interface, utilizing WSDL and REST, to expose data sources. By utilizing mediation technology, data services will be able to integrate data exposed via different services. Hosting VPH-Share services follows the virtual appliance approch enabling the hosting of services in the Cloud. Client applications usually access data services by submitting an SQL query, or, in the case of semantically

annotated data sources, a SPARQL query, and download the query results in respective formats (e.g. WebRowSet or RDF triples). The data service framework internally utilizes established data access and integration technologies capabilities including OGSA-DAI [8] and OGSA-DQP [9].

The hosting of data services relies on the concept of virtual appliances. A virtual appliance can be defined as a software package pre-installed on a virtual machine image to enable provisioning of the software in the Cloud. The VPH-Share Cloud infrastructure will enable on-demand hosting of data sources exposed as services and provided via virtual appliances in the Cloud.

4 VPH Semantics

The VPH-Share semantics layer aims at providing 'knowledge level' functionality to the stakeholder by establishing an abstraction over the lower-level compute and data services.

VPH-Share will provide facilities for assisting users in selecting suitable ontologies and annotating datasources with them. Informed by these annotations VPH-Share will provide means for exposing and integrating distributed datasets exploiting linked data principles [15]. In particular, supported by this technology the project shall support accessing the underlying information through different semantic views and combining these different view for carrying out global analysis. Similarly, VPH-Share will provide support for annotating computational services so as to exploit these annotations in order to better assist data analysts in the discovery of applicable analysis services, as well as to help composing and invoking them. In this respect, the project shall leverage linked services technologies notably their integration with linked data as a processing infrastructure [16].

Finally, supported by the ability to integrate and process distributed data, the project shall devise a number of data inference services. These services will leverage domain knowledge, data mining, and machine learning technologies to analyse the wealth of information captured in order infer and estimate additional information, thus allowing practitioners to reach previously unattainable insights which would presumably lead to further and better informed decisions.

5 Related Work

Building an infrastructure for modelling and managing biomedical information has been addressed by multiple projects. The @neurIST project dealt with supporting the research and treatment of cerebral aneurysms. An advanced service-oriented IT infrastructure for the management of all processes linked to research, diagnosis, and treatment development for complex and multi-factorial diseases has been developed.

Another project in the domain of VPH, called Health-e-Child, creates an information modelling methodology based around three complementary concepts: data, metadata, and semantics. The goal is to give clinicans a comprehensive view

of a child's health by integrating biomedical data, information and knowledge. The utilized data spans from imaging to genetic to clinical and epidemiological.

The caCORE infrastructure [11], developed by the National Cancer Institute (NCI), United States provides tools for the development of interoperable information management systems for data sharing and is particularly focused on biological data in the cancer domain. Additional projects, relying on the model-driven software architecture of caCORE for managing biomedical research information haven been started (caGrid [12], CaBIG [13]).

The PhysiomeSpace [14] is a digital library service for biomedical data and has been developed in the LHDL project. PhysiomeSpace provides services for sharing biomedical data and models with a mixed free/pay-per-use business model that should ensure long term sustainability.

6 Conclusion

The VPH-Share project is part of the VPH initiative of the European Commission with the goal of providing a systematic framework for understanding physiological processes in the human body in terms of anatomical structure and biophysical mechanisms at multiple length and time scales. The project will provide a systematic framework for the understanding of physiological processes in the human body.

The VPH-Share project introduces the concept of a patient avatar which can be defined as a coherent digital representation of each patient including information relevant to different contexts. For managing data the project relies on the DIKW pyramid describing the path from data, information, and knowledge, to wisdom. A flexible, semantically enhanced, data management platform will be created supporting this approach and relying on the concept of data services to enable the vision of patient avatars.

The infrastructure will be developed based on the requirements of four flagship workflows (@neurIST, euHeart, VPHOP, and Virolab). These workflows serve as the empirical basis and benchmarks for the support structure to be developed.

The VPH-Share project will provide a Cloud infrastructure facilitating on-demand access to data and compute resources. On top of the Cloud infrastructure, VPH-Share services including data access, data mediation, semantic, and compute services, as well as workflows, will be hosted on demand.

References

1. ARGOS: Transatlantic Observatory for Meeting Global Health Policy Challenges through ICT-Enabled Solutions (2011), `http://argos.eurorec.org/`
2. Hunter, P., Coveney, P., de Bono, B., Diaz, V., Fenner, J., Frangi, A., Harris, P., Hose, R., Kohl, P., Lawford, P., McCormack, K., Mendes, M., Omholt, S., Quarteroni, A., Skår, J., Tegner, J., Thomas, S., Tollis, I., Tsamardinos, I., van Beek, J., Viceconti, M.: A vision and strategy for the virtual physiological human in 2010 and beyond. Phil. Trans. Royal Society A, 2595–2614 (2010)

3. Dunlop, R., Arbona, A., Rajasekaran, H., Iacono, L.L., Fingberg, J., Summers, P., Benkner, S., Engelbrecht, G., Chiarini, A., Friedrich, C.M., Moore, B., Bijlenga, P., Iavindrasana, J., Hose, R.D., Frangi, A.F.: @neurIST - Chronic Disease Management through Integration of Heterogeneous Data and Computer-interpretable Guideline Services. In: Proceedings of Healthgrid (2008)
4. Franklin, M., Halevy, A., Maier, D.: From Databases to Dataspaces: A new abstraction for information management. ACM SIGMOD (2005)
5. Benkner, S., Engelbrecht, G., Koehler, M., Woehrer, A.: Virtualizing scientific applications and data sources as grid services. In: Cao, J. (ed.) Cyberinfrastructure Technologies and Applications. Nova Science Publishers (2009)
6. Koehler, M., Benkner, S.: VCE - A Versatile Cloud Environment for Scientific Applications. In: The Seventh International Conference on Autonomic and Autonomous Systems, ICAS (2011)
7. Benkner, S., Arbona, A., Berti, G., Chiarini, A., Dunlop, R., Engelbrecht, G., Frangi, A.F., Friedrich, C.M., Hanser, S., Hasselmeyer, P., Hose, R.D., Iavindrasana, J., Koehler, M., Iacono, L.L., Lonsdale, G., Meyer, R., Moore, B., Rajasekaran, H., Summers, P.E., Woehrer, A., Wood, S.: @neurist: Infrastructure for advanced disease management through integration of heterogeneous data, computing, and complex processing services. IEEE Transactions on Information Technology in Biomedicine 14(6), 1365–1377 (2010)
8. Antonioletti, M., Atkinson, M., Baxter, R., Borley, A., Hong, C., Neil, P., Collins, B., Hardman, N., Hume, A.C., Knox, A., Jackson, M., Krause, A., Laws, S., Magowan, J., Paton, N.W., Pearson, D., Sugden, T., Watson, P., Westhead, M.: The design and implementation of grid database services in ogsa-dai: Research articles. Concurrency and Computation: Practice and Experience 17(2-4), 357–376 (2005)
9. Alpdemir, M.N., Mukherjee, A., Gounaris, A., Paton, N.W., Watson, P., Fernandes, A.A.A., Fitzgerald, D.J.: OGSA-DQP: A Service for Distributed Querying on the Grid. In: Bertino, E., Christodoulakis, S., Plexousakis, D., Christophides, V., Koubarakis, M., Böhm, K. (eds.) EDBT 2004. LNCS, vol. 2992, pp. 858–861. Springer, Heidelberg (2004)
10. Branson, A., Hauer, T., McClatchey, R., Rogulin, D., Shamdasani, J.: A Data Model for Integrating Heterogeneous Medical Data in the Health-e-Child Project. In: Proceedings of HealthGrid (2008)
11. Komatsoulis, G.A., Warze, D.B., Hartel, F.W.: caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability. Journal of Biomedical Informatics 41(1), 106–123 (2008)
12. Oster, S., Langella, S., Hastings, S., Ervin, D.: caGrid 1.0: An Enterprise Grid Infrastructure for Biomedical Research. Journal of the American Medical Informatics Association 15(2), 138–149 (2008)
13. Buetow, K.H., Niederhuber, J.: Infrastructure For A Learning Health Care System: CaBIG. Health Affairs 28(3), 923–924 (2009)
14. Testi, D., Quadrani, P., Viceconti, M.: PhysiomeSpace: digital library service for biomedical data. Phil. Trans. R. Soc. 368(1921), 2853–2861 (2010)
15. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS) (2009)
16. Pedrinaci, C., Domingue, J.: Toward the Next Wave of Services: Linked Services for the Web of Data. Journal of Universal Computer Science 16(13), 1694–1719 (2010)