EUDAT: A Collaborative Data Infrastructure Supporting the Virtual Physiological Human Initiative

Ali N. Haidar[1], Stefan J. Zasada[1], Damien Lecarpentier[2], Peter Wittenburg[3] and Peter V. Coveney[1]
[1] University College London, UK
[2]CSC – the Finnish IT Centre for Science, Finland
[3]Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

1. Introduction

European scientific and research communities from a wide range of disciplines, including the Virtual Physiological Human (VPH) community [1], are faced with increasingly large amounts of valuable data that stem from existing sources such as patient data and new sources, such as new sensors and scientific instruments used in analyses, experiments and observations as well as growing volumes of data from simulations and the digitisation of resources such as libraries [2]. The accelerated proliferation of data has created a need to tackle the specific challenges of data management, and to ensure a coherent approach to research data access and preservation. Moreover, the current data landscape in Europe is still fragmented, with most initiatives addressing the needs of a specific discipline or community. For example, the European Bioinformatics Institute (EBI) [3] and the European Space Agency (ESA) [4] generate and manage massive volumes of data such as genome, protein sequences, gens, enzymes and spatial imaging relevant to biomedical and environmental scientists' communities. This has resulted in increasing diversity with respect to data architectures, organizations, formats and semantics. Issues related to integration and interoperability of existing data infrastructures are a growing concern. Rising costs due to the explosion of data are also threatening the financial viability of those infrastructures.

We describe the European Data Infrastructure project (EUDAT) [2] co-funded by the European Commission's Framework Programme 7 and comprised of 25 European partners including data centres, technology providers, research communities and funding agencies from 13 countries. EUDAT aims to contribute to the production of a Collaborative Data Infrastructure (CDI) driven by researchers' needs and focuses on the data management. Research communities from different disciplines use different types of data and organise it in different ways, but they also share basic service requirements. For example, long-term data archives for integrity and authenticity control in many research communities, and a shared demand for data federation and services enabling discovery, access, data mining, integration and curation. These common requirements mean that it is desirable to establish generic pan-European services designed to support multiple communities, as part of a collaborative framework. Building this common layer of generic and cross-disciplinary data services is precisely the focus of EUDAT project. VPH community [5] is represented in EUDAT via the VPH-Network of Excellence (VPH-NoE) to communicate the needs of the wider biomedical computing communities, including ECRIN [6], ELIXIR [7], p-medicine [10] and EuroBioimaging [8].

2. Overview of EUDAT and how the VPH Community will benefit from it

The VPH Initiative heavily relies on heterogeneous patient data in order to perform patient specific modelling and simulations. Patient specific refers to the tailoring of medical treatments based on the characteristics of an individual patient. Decision support systems based on patient specific computer simulation using real patient data collected from various hospitals on grid infrastructures hold the potential of revolutionising the way clinicians plan courses of treatments for patients. Patient data consists of imaging data in DICOM format (MRI, CT, PET), micro-photos in the JPEG format, treatment data in Excel sheets format, genetic data, simulation model data and other omics data related to a disease such as Cancer [9,10] and HIV [11]. Many VPH projects share the same data life cycle, shown in Figure 1, which involves data acquisition from several hospitals, data collection, transfer of the data to some repositories*, a secure integrated data storage and sharing environment* and *a security framework for ensuring data protection*. EUDAT focuses on the latter two items.

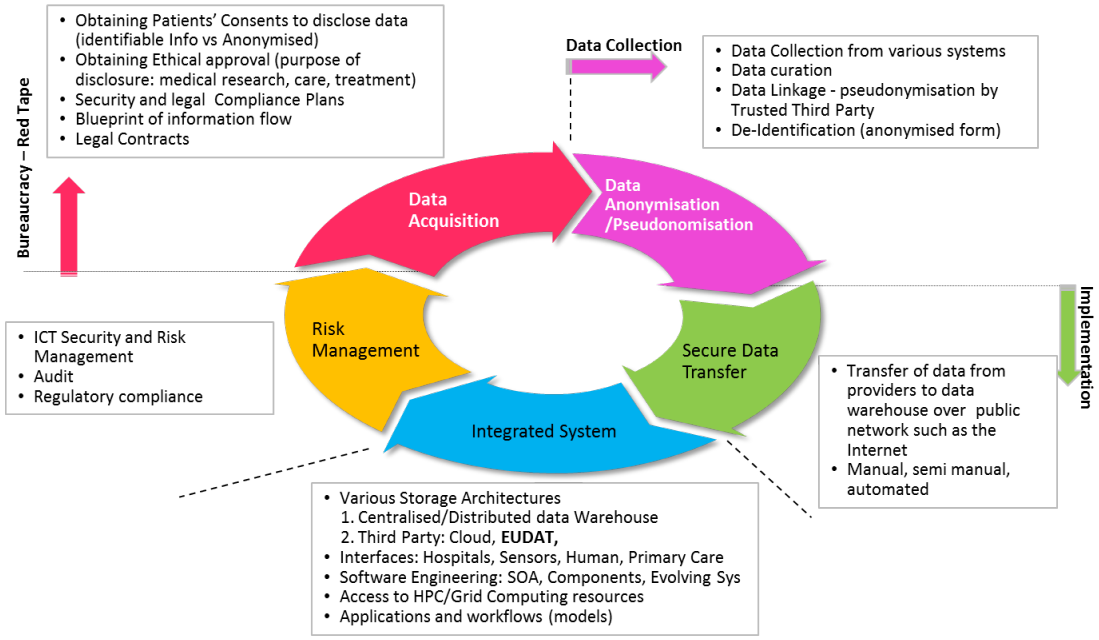

Figure 1. VPH Data lifecycle: using EUDAT for VPH data storage, archiving and sharing

To illustrate how EUDAT can support VPH research projects, we present the following examples done as part of EUDAT requirement gathering process for the VPH community.

ContraCancrum [9,12] was a data infrastructure project running until August 2011 and is now being followed up by p-medicine [10]. ContraCancrum involves data aggregation from a number of hospitals into one central repository to make it accessible to researchers. The size of the data collected as a demonstrator exceeded 20 Terabyte and required dedicated personnel to manage it. The software setup used to do the aggregation is called IMENSE [12] and allows scientists to run workflows using the stored patient data on remote grid and HPC resources. P-medicine will continue the aggregation efforts based on the IMENSE software, but will step away from a central solution and work towards a federation of data hubs in various countries. The key parts relevant to EUDAT in these projects are the data storage and the security framework to protect the data. These two projects heavily invested in building and deploying these parts in-house.

The benefits associated with creating a Collaborative Data Infrastructure, in which research communities can rely on a set of common data services will result in better exploitation of synergies. The collaborative data infrastructure put in place by EUDAT will help to support the data infrastructures currently being used and developed by scientific communities by offering them an infrastructure on which they can rely for their more generic data needs. This will allow research communities to focus a greater part of their effort and investment on services that are specific to their discipline. The EUDAT data infrastructure will also provide individual researchers, smaller communities, and projects lacking tailored data management solutions with access to sophisticated shared services*, removing the need for large-scale investment in infrastructure development.*

Research communities such as the VPH can then build on these generic data services to provide rich, community specific analysis platforms. The fact that many EUDAT partners are also large HPC centres participating in PRACE should make it easy for VPH researchers to collocate their data with high performance computing. This is particularly useful for many multi-scale VPH research projects [5] (ContraCancrum, euHeart, VPH-Share, preDict) where workflows are central functionalities. Currently, all EUDAT storage facilities are to be located adjacent to large supercomputing centres (such as Tier-1 and Tier-0 site members of PRACE), to facilitate data exchange with computers that themselves produce large quantities of data.

3. EUDAT progress

To build a sustainable data infrastructure upon which common services can be deployed for use by diverse communities, a comprehensive approach is required, including several activity strands.

EUDAT is currently investigating user requirements, starting with research communities in linguistics (CLARIN) [13], earth sciences (EPOS) [14], climate sciences (ENES) [15], environmental sciences (LIFEWATCH), and biological and medical sciences (VPH), which have been allocated project resources to help specify their requirements and co-design related services via interviews. This investigation will be extended to additional communities in 2012. A second activity strand concerns the appraisal of technologies and service candidates, which involves identifying, designing and constructing appropriate services, using existing solutions where possible. This will also help in identifying the gaps that should be addressed by EUDAT. The third activity strand involves primarily the data centres and deals with the operation of the collaborative infrastructure, particularly the provisioning of secure, reliable (generic) services in a production environment, with interfaces for cross-site and cross-community operation. The operation of the infrastructure should provide full life cycle data management services, ensuring the authenticity, integrity, retention and preservation of data, especially those marked for long-term archiving.

4. Conclusion

EUDAT aims to provide a persistent, robust data infrastructure and services for the VPH community, as well as to others, that fits well with requirements of the VPH Initiative. Building the Collaborative Data Infrastructure will not be a trivial task. It will require active collaboration between all actors, and in particular between the communities involved in designing specific services and the data centres willing to provide generic solutions. We must also plan, from the very beginning, the evolution and sustainability of the infrastructure. Among other things, this implies early definition of future partnership and business models for adopting, supporting and sustaining common services developed for, and partly operated by, the different research communities. To achieve this, we first need to show that our service approach is feasible; therefore the design and deployment of early services will be critical for the success of the project. Data reuse in an open data infrastructure scenario also implies that data creators, managers and users no longer know each other: they are acting anonymously, but nevertheless must rely on each other's quality of work. Thus new mechanisms are also necessary to establish trust between all stakeholders.

Acknowledgments: This work is supported by EUDAT [2] (FP7-2007-2013) and Virtual Physiological Human Network of Excellence VPH-NoE [1] (FP7-2007-IST-223920).

References

[1] The Virtual Physiological Human Network of Excellence http://www.vph-noe.eu/

[2] European Data Infrastructure (EUDAT) www.eudat.eu

[3] The European Bioinformatics Institute (EBI) http://www.ebi.ac.uk/

[4] The European Space Agency http://www.esa.int/esaCP/index.html

[5] List of VPH projects: http://www.vph-noe.eu/vph-projects

[6] European Clinical Research Infrastructures Network (ECRIN) http://www.ecrin.org/

[7] European Life Sciences Infrastructure For Biological Information (ELIXIR), http://www.elixir-europe.org/

[8] EuroBioimaging http://www.eurobioimaging.eu/

[9] Clinically Oriented Translational Cancel Multilevel Modelling (ContraCancrum) http://www.contracancrum.eu/

[10] Personalised Medicine (P-medicine): http://p-medicine.eu/

[11] ViroLab: http://www.virolab.org

[12] S. J. Zasada , T. Wang , A. Haidar , E. Liu, B. Graf, G. Clapworthy, S. Manos, P. V. Coveney, "IMENSE: An e-Infrastructure Environment for Patient Specific Multiscale Modelling and Treatment", Journal of Computational Science, (2011), DOI: 10.1016/j.jocs.2011.07.001, http://www.sciencedirect.com/science/article/pii/S1877750311000639

[13] Common Language Resources and Technology Infrastructure(CLARIN) http://www.clarin.eu/

[14] The European Plate Observing System (EPOS) http://www.epos-eu.org

[15] Infrastructure for the European Network for Earth System Modelling http://is.enes.org