Services Design in a Collaborative Network for Multidisciplinary Research Projects

Maria Krestyaninova[1,2] and Yulia Tammisto[3],

[1] Institute for Molecular Medicine Finland, FIMM, University of Helsinki, Biomedicum Helsinki 2U, 00014 Helsinki, Finland
{Masha}@simbioms.org

[2] Uniquer Sarl, 12 rue de la Mercerie, Lausanne, 1003, Switzerland

[3] Aalto University School of Economics, Runeberginkatu 14, 00100 Helsinki, Finland
{Yulia.Tammisto}@aalto.fi

Abstract. Providing information management services for multi-disciplinary research projects presents scientific, technological and communication challenges: constant change of themes and objects of scientific studies, entangled privacy and ownership requirements along with rapidly evolving methods of analysis and ways to look at data call for a highly dynamic communication and data service. We will present a case study of a collaborative network (simbioms.org). SIMBioMS, founded in 2005, develops open source software and provides data management services in biomedical research through four academic organisations and one company.

Keywords: multidisciplinary project, research consortia, open data, data management, service design

1 Introduction

Post-genomic era in life sciences has arrived with an unprecedented number and scale of multidisciplinary collaborative projects [1]. It is no longer sufficient to look only at one type of data in order to report a discovery in environmental or biomedical research. Translation to practical applications is imperative [2,3]. Complex studies utilizing several technological platforms to assay thousands of biological samples have become a norm. Technological and analytical advances are reinforcing this trend [4]. However, work and expertise required for such multidisciplinary research projects cannot be delivered by a single organization. Collaborative networks and

research consortia are being organized and compete with each other at a remarkable rate. Communicating ideas, study designs, complex data sets and information regarding data structures is the key challenge for the success of a collaborative project [5].

1.1 Central bioinformatics services and new challenges

Originally, IT services in life sciences research were dedicated to collecting and preserving the data for posterity in a uniform fashion (often as a compulsory exercise precluding a publication in a peer-reviewed journal), and to making it available to the worldwide scientific community. This was delivered by large centralized archives, funded by governments [6,7].
However archives were unable to support new, collective fashion of carrying out studies by research consortia with large number of participants. Such studies often deal with data that often cannot be released to general public (population–wide human genomics), their participants need to exchange data prior to discovery (at the stage of study design), there are complex requirements regarding authorship and administration of a study. Combined, these factors have led to formation of a new type of services [8, 9]. Since 2004-2005 there has been a wave of open source and proprietary software and services initiatives that undertook a mission of supporting complex cross-organisational communications that were not provided for by large public archives in life sciences (Fig. 1).

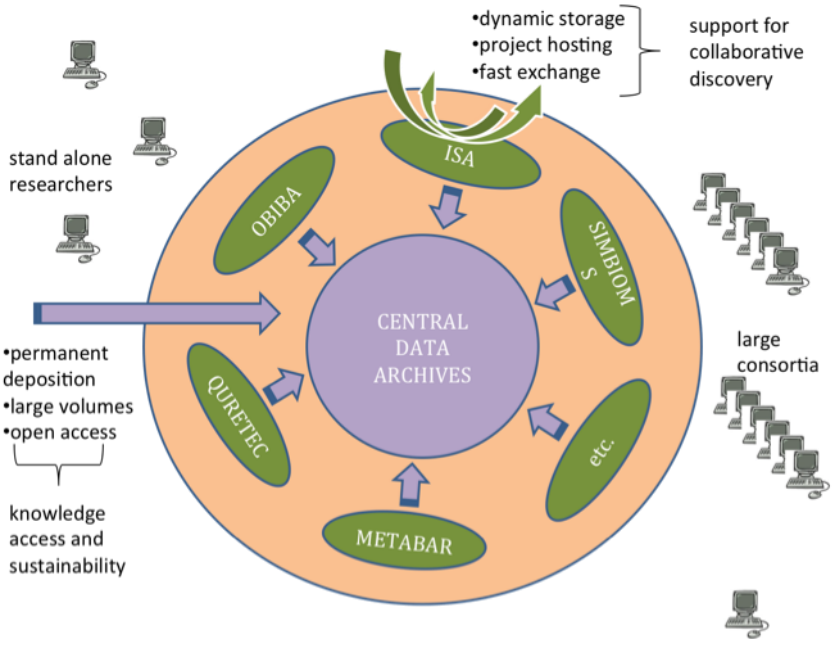

Fig. 1. Central archives (purple) collect all the data and make it available for everyone, open access; main mission: knowledge accumulation and dissemination for the entire scientific community. Dynamic collaborative IT platforms (green) facilitate efficient dataflow and other communications within a group of collaborators in the context of a specific study; main mission: provide means of communicating analysis-related information throughout a project; assist with data deposition upon completion of a project.

1.2 Solutions for collaborative discovery

Currently, such service providers do not serve only a specific community, as it was in the beginning. Generic use-cases, communication flow and reporting requirements in various fields of research have been formulated [10,11] and implemented in numerous instances. Often the same provider would have customers in ecology, human genetics, pharmaceutical and biomolecular fields and would implement services simply by configuring the same software. Demand for consortium services continues to grow, and, as service providers gain domain-specific expertise, competition between IT players also increases.

If one was to generalise the implications of the multidisciplinary collaboration trend on the evolution of research data management services, one could name three fast-developing directions:

1) means for fine-tuning of access rights for data owners: data can no longer be categorized into completely private or completely open
2) instruments for handling heterogeneity of data, metadata and the variety of standards
3) accessibility and suitability of the source code used in services construction

Research groups nowadays deploy tools that assist in structuring and describing their data in a way, which is most useful for the scope of their study and for the format of their communication with the collaborators. The underlying architecture and semantics of those tools is compatible with international data formatting standards [12].

1.3 Distributed software development and services provision

Whilst the majority of service providers for research consortia have put effort into creating generic software and, therefore, services that they provide consist of installing and configuring their own software (i.e., service and software are inseparable), we pursued a scenario in which we created a collaborative IT network with four academic institutions and one company as contributors (simbioms.org). The motivation behind such a set-up lays in avoiding a long-term dependency on a specific architecture or software platform. We aimed to develop a mechanism for building partnerships between several software suppliers and a service provider and thus to respond promptly to newly emerging needs and rapidly expand across knowledge domains without loss in sustainability or quality. One of the major challenges that we faced as a network was the absence of services design culture in

bioinformatics: creating/configuring software and developing a service were one and the same.

2 Results

The SIMBioMS collaborative network, Systems for Information Management in BioMedical Studies, was created in 2005 when IT specialists from two academic institutions – Institute of Mathematics and Computer Science (Riga, Latvia) and European Bioinformatics Institute (Hinxton, UK) - formed a working group to provide data management services for an EU-funded large research project [13]. Since 2007, software engineers, data managers and system administrators from Karolinska Institutet (Stockholm, Sweden), Institute for Molecular Medicine Finland (Helsinki, Finland) and Uniquer (Lausanne, Switzerland) have joined the network. The establishment of SIMBioMS as a virtual enterprise included:

- web visibility: domain, website
- central code management: CVS (Concurrent Versions System), git (Distributed version control and source code management system)
- standard operation procedures (SOPs) for code release, testing and deployment for services
- single-point user support: email, ticketing and other services
- internal communications: wiki, email lists, meetings and teleconferences

2.1 Software and services.

The network so far has produced 5 modules of software for data management and for administration design of complex meta-studies [14] and provided services in more than 10 EU and national collaborative projects (see Table 1).

Table 1. List of projects in which SIMBioMS has provided data exchange services. Contribution to various projects on strategy and design (ELIXIR, BBMRI, P3G and TaraOceans) aren't included.

Projects	N partners	Description	N samples/data files
ENGAGE	25	Genome-wide association studies	>100k
SUMMIT	24	Genetic basis of diabetic complications	65k
SIROCCO	27	siRNA research consortium	20k
Biomedinfra.fi	4	Finnish national biobank network	70k
sail.simbioms.org	14+	Sample availability on-line resource	180k
MOLPAGE	18	Molecular phenotyping, 11 HT platforms, multiple tissues	25k
MuTHER	4	SNP, RNA, methylation in multiple tissues	1k
EGG	16	Early growth genetics	5k
CAGEKID	14	Cancer genomics of kidney	3k
ENGAGE	25	GWAS	>100k
BIOBANQUES	70	French national biobank network	1mln

The software is modularized and customizable. It is compatible with the major public archives and integrates well with other software products available in the same niche. Tutorials, guides, and documentation for installation and use are provided along with a service provision.
Services for each of the projects have been implemented using 2 or 3 modules depending on project's communication flow. For each project there is a number of configurations (up to 25 per project per module) and services blueprints. Requirements analysis, design of the services and of the release cycle were done jointly by a service providing partner and a software developing partner.

2.2 Partnership formation.

The network has a procedure for introducing new partners and additionally it has experience in providing services jointly with external IT players. The longevity of a partnership depends on whether it is happening within a specific project or line of funding, or beyond it. In the latter case a group is considered a part of the network, while in a project-based partnership – an external collaborator. Licensing agreements (software must be open source), code of conduct, communication set-up with the network have been put in place in order to clarify the terms of membership in the network. Issues regarding acknowledgement, IP ownership and, generally, ethical and legal settings of the network remain work in progress.

2.3 Collaborative framework for IT services provision

Two types of intellectual value generated by the SIMBioMS network have been established: configurations (metadata structures) and services blueprints. Over 7 years, there have been over 100 configurations and over 10 blueprints created by the partners. Software developing partners capitalized on the former, while service providers benefitted from the latter. The two values are essentially the currency within the virtual enterprise. The communications and work within the network are aimed at increasing both types of intellectual value. Collaboration with external partners, e.g. large public archives, is anchored to blueprints and configurations as well. The larger the study-specific collection of use-cases and data structures the more complex the design of the tools (blueprints and configurations) for data collection, re-annotation and standardization, as well as the design of the interface that serves the data to the community.

3 Discussion

Multidisciplinary and cross-organisational nature of research calls for more advanced communication systems. The notion of open data is no longer as clearly defined as it used to be: in the process of discovery one does need to open data consecutively, first

to collaborators, then to publishers and then to the general public. Research collaborations urgently require some dedicated virtual environment that can ease the administrative and communication burden in large collaborative projects. It is often hard to pre-define a communication structure due to numerous legal, ethical, psychological, technological and intellectual challenges. We argue that not only creation, but rather succession and sustainability of research consortia and results it produces collectively can be helped and enhanced by a well-designed communication and data exchange platform. Such a platform would allow:
-to share data selectively, and leave access control in the hands of those who are responsible for extracting value from data,
- to declare an intended study in standardized terms that later on can be tracked from the resulting publication or patent.
The main requirements for such an IT platform for collaborative discovery are data semantics services (translation between terminology, assistance in tagging, normalizations) and the means to fine-tune data access.

Designing software for a wide variety of constantly evolving research themes is a massive task, unless data generators are forced to convert data into a universal standard. The services design approach could become a suitable alternative: capturing service blueprints and corresponding configurations along different studies can help to develop more efficientinformation management solutions for large biomedical studies. However, it requires a sustainable partnership between service providers and open source software suppliers.

Undoubtedly, it is possible to produce software and design services at the same time. In bioinformatics the two activities are not usually separated and users participating in the design of services are often unaware that they contribute significantly to the software design. The more complex the dynamics of collaboration is, the more efforts are required for crafting an IT service that mimics this complex communication.

When talking about the communication infrastructure for multidisciplinary collaboration, the focus shifts from software production to creating and capturing metadata structure and services blueprints. Information about interactions between the players of research collaborations represents the intellectual value that is equally significant to technological value of the software. Intellectual value is gathered collectively by scientists exchanging the data, software engineers and service designers.

Establishing simple ways to exchange designs among the network members would enable the players to sustain collaborative IT networks beyond one specific project, and that would yield a sustainable IT infrastructure. It remains to be seen what legal and ethical instruments can be put in place in order to strengthen partnerships required for designing bioinformatics services or even to develop a culture for exchanging of designs within the biomedical research community more effectively.

In order to provide reliable and efficient services IT groups themselves will have to gain a deeper understanding of dynamics and nature of joint work. Therefore, longevity and sustainability of collaborative links in bioinformatics services and software development are crucial for complete and comprehensive collection of requirements from the entire research ecosystem.

Acknowledgments. This research was supported through funds from The European Community's Seventh Framework Programme (FP7/2007-2013), ENGAGE Consortium, grant agreement HEALTH-F4-2007-201413. The authors would like to thank everyone who participated in, contributed to and funded SIMBioMS work in the course of 7 years. The full list of people would be over 100 names and over 60 institutions and organizations. Special thanks to those who delivered services and software under SIMBioMS umbrella: Sudeshna Guha Neogi (NIHR), Teemu Perheentupa (FIMM), Jani Heikkinen (FIMM), Juha Muilu (FIMM), Joern Dietrich (Uniquer Sarl), Natalja Kurbatova (EMBL-EBI), Julio Fernandez-Banet (EMBL-EBI), Janna Hastings (EMBL-EBI), Mikhail Gostev (EMBL-EBI), Stathis Kanterakis (EMBL-EBI), Sandra Ose (IMCS), Juris Viksna (IMCS), Andris Zarins (IMCS), Edgars Celms (IMCS), Russell Vincent (Uniquer Sarl), Inga Prokopenko (OCDEM, WTCHG), Huei-Yi Shen (FIMM), Ola Spjuth (KI). We would also like personally thank Ugis Sarkans (EMBL-EBI) who beside being a valuable member of SIMBioMS network, has significantly contributed to the development of this paper by discussing our approach, suggesting ideas and commenting on the text

References

1. Mesko B, et al: The triad of success in personalised medicine: pharmacogenomics, biotechnology and regulatory issues from a Central European perspective. N Biotechnol. 2012 Mar 10
2. Prokopenko, I. et al: Variants in MTNR1B influence fasting glucose levels. Nature Genetics (2009), 41:77-81.
3. Karsenti E, et al: A Holistic Approach to Marine Eco-Systems Biology.PLoS Biol 9(10): e1001177
4. McConnell P, et al. The cancer translational research informatics platform. BMC Med Inform Decis Mak (2008) 24;8:60.
5. Anderson, N.R. et al.: Issues in Biomedical Research Data Management and Analysis: Needs and Barriers. J Am Med Inform Assoc. (2007), 14(4):478–488.
6. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov
7. European Bioinformatics Institute, http://www.ebi.ac.uk
8. Rocca-Serra, P. et al.: ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics (2010), 26(18):2354-2356.
9. Smedley, D., et al.: BioMart – biological Queries made easy. BMC genomics (2009), 10:22.
10. Smith, B. et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. (2007), 25:1251–1255.
11. Quackenbush, J.: Standardizing the standards. Mol.Syst. Biol. (2006), 2: 2006.0010.
12. Sansone SA., et al.: Towards interoperable bioscience data. Nat Genet. 2012 Jan 27;44(2):121-6
13. Nicholson G., et al: A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. PLoS Genet (2011). Sep;7(9):e1002270
14. Krestyaninova, M. et al.: A system for Information Management in BioMedical Studies-SIMBioMS. Bioinformatics (2009), 25(20):2768-2769.