Assembling an IT Infrastructure in Data Intensive Collaborative Projects in the Life Sciences

Stathis KANTERAKIS*
EMBL-EBI, European Molecular Biology Laboratory, Wellcome Trust Genome Campus
Hinxton, Cambridge, CB10 1SD, UK.

and

Maria KRESTYANINOVA
FIMM Institute for Molecular Medicine Finland, Helsinki University
Helsinki, FI-00014, Finland.
*Corresponding author (kanterae@ebi.ac.uk)

ABSTRACT

Wide adoption of novel analytical technologies in genetics and advancements in data analysis that allowed to pool data from several data collections, have created a need for web-based data integration and harmonisation services [1-4]. Research consortia are seeking for IT platforms that can enhance the communication between biologists, statisticians, geneticists and clinicians and through this reduce the burden of data management and administration tasks. Design and implementation of a communication and data exchange platform is crucial for the efficient resource allocation and fruitful data analysis in large research consortia [5,6,7].

So far, most information systems for molecular, genetic, clinical and life-style data have been designed as long-term repositories with strict formats and requirements [8, 9]. The primary mission of such repositories is to preserve data for posterity in the most uniform fashion and make it available to researchers worldwide. Demands for short-to-medium term data deposition and assistance in data handling during the creative, discovery phase of a research project have not yet been addressed. Project-specific data management platforms scalable to population-size datasets and flexible enough to deal with very diverse biological, medical and life-style data are vital for researchers, since they speed up data exchange, annotation and integration for the context of a specific study [10, 11].

We present a novel framework for data intensive communications in large collaborative projects. First, we analyse current trends in collaborative knowledge generation, online information exchange and peer design, and then apply these principles in the life sciences, in the context of large collaborative studies. We come up with common use cases and corresponding software modules to enable such usage, and propose a vision for the design principles that should permeate collaborative biomedical software in the future.

Keywords: collaborative research, open design, omics, information management, biomedical studies, knowledge creation

1. INTRODUCTION

"On the one hand", claims writer Steward Brand in the late 1980s, "information wants to be expensive, because it's so valuable. The right information in the right place just changes your life. On the other hand, information wants to be free, because the cost of getting it out is getting lower and lower all the time"[12]. The latter is more relevant than ever, a quarter of a century later. We live in an era of abundant information. Since the sequencing of the first human genome, a map of our genetic "blueprint", and the advent of high-throughput technologies in biomedical research, information has exploded in the life sciences. This information is expensive; a stand-alone research laboratory can no longer afford the machinery and expertise to generate it. The cost of sequencing a single sample, while steadily declining, adds up as we try to detect increasingly rare genetic variants using larger cohorts. Information is also free; a push for open-data and open-science in the recent years, and the adoption of such standards from scientific journals as a requirement for publication means there is a lot of "free" information out there. Data volume and complexity are soaring faster than any analysis can hope to tackle. How do we distribute the financial burden of data that is costly to generate? How do we handle the increasing volume and complexity of available data that is sprinting further and further from our capacity to examine it? Furthermore, what does it take to convert information to knowledge, or to translate scientific findings to clinically relevant results? [13]

Open innovation

The revolution of open innovation in business, which arrived a decade ago, is only recently appearing in the life sciences. Open innovation refers to the model of allowing ideas to flow freely outside a firm's boundaries, becoming licensing deals or spin-offs, as well as from the outside into a firms boundary, becoming part of its IP portfolio and core development efforts [14]. What made this possible? The availability of private Venture Capital funding allowed ideas to materialise independently of a firm's primary focus. If the R&D department was not willing to fund a project, visionaries could look outside for their own funding. Good ideas got a chance of survival, while bad ideas were quickly replaced by better ones from the

market. The dawn of such funding opportunities in science, notably with SciFlies.org [15], a website which lists and promotes peer-reviewed projects for funding directly from the online community, means science is becoming in part more independent. Think tanks such as the Open Science Working Group advocate open publication (free of charge to the reader) as well as open deposition of scientific data to public repositories in a useable form [16]. The European Bioinformatics Institute (EMBL-EBI) is in the forefront of the open data effort, creating standards such as minimal information requirements, to make published data interpretable and usable, without the need to contact the primary source [17,18,19]. It also offers access to state-of-the-art data archives, free of charge [20,21].

Biological science is becoming more distributed. Because of the cost and complexity of modern "omics" research [22], it is unlikely that all resources and expertise to tackle a biological problem will exist under a single roof. For that reason, pooled research, such as in forming consortia, has become popular in modern biological research [1-4]. What technically makes data pooling and integration possible is consistent annotation. For that reason, terminologies, nomenclatures, ontologies and other types of controlled descriptors are being developed and maintained centrally, through curation and community involvement. When several datasets are merged for the sake of meta-analysis, this is often done through the application of a standard format (e.g. MAGETAB [23], ISATAB [24]) or simply through re-annotation and transformation of data to a common lexical denominator (DataShaper [11], SAIL [25], or tools for semantic tagging). Many such tools empower researchers with the means to carry out annotation tasks in their local laboratories, thus distributing the burden of curation from central repositories to the data source and local expertise. The differences, benefits and possible shortcomings of the open innovation model in biological research are outlined in Table 1.

Table 1. Comparison between open, distributed research and the traditional closed discovery model from the perspective of a single research institution.

	Closed model	**Open model**	**Benefits of the open model**	**Drawbacks of the open model**
Funding	Individual public funding	As part of a consortium / directly from the community	Increased accountability, peer pressure	Less competition, motivation
Scientific community involvement	Research behind "closed doors"	Intermediate results immediately available to the community	More discussion, feedback, cross-disciplinary communication	Proper attribution of credit for discovery, security, ethical issues
Curation	At central repositories, if at all	By local experts, at central repositories and community	Higher quality research	Increased scrutiny, especially against challenging the status quo
Visibility	Not a high priority	Standards enforced, tools for better contextualization, cross-linking of information	Objectivity, re-use of data and resources	More effort upfront
Value	Realized by a single laboratory	Realized by the whole community	Better return for funders' investment, better knowledge generation potential	Increased stress, push for publication rather than knowledge generation

Finally, promising community authorship attempts, such as wikigenes.org, which is built around the creed of proper author attribution, hint at the way scientific knowledge might be organised in the future. It is not a stretch of imagination to think of an era where the journal publication will be a from of rhetoric art, serving its own purpose, while scientific facts will be organised in a highly inter-connected, immediately and openly accessible medium and in a useful manner. This would not only save time and energy, but also propel scientific discovery. It would be analogous to a family collaboratively solving a gigantic puzzle, placing each piece straight into its proper place during each move.

A central point in the discussion about open data has always been the fear of its "incompetent use", who should be concerned about it, and which data is to be released: raw data, processed data or results? Interestingly, these three "levels" correspond to the three stages of knowledge creation, from raw material to information to knowledge. Scientists are often happy to disclose the latter two levels but reluctant to do so with the first, be it in fear of being swept by competition, uneasiness in releasing "garbage" data, infringement of legal or ethical regulations, such as to research subjects, or concern regarding proper acknowledgement for generating the data. This, by definition, cripples the ability of further information and knowledge creation since it forces data to be adopted in the view that the generator intended. This is understandable, in part, due to disincentives in place to promote such behaviour: scientists being judged by volume of publications and not necessarily quality, the inability to externalize costs related to production and enforcement of intellectual property rights, consent by research subjects to only particular kinds of research (e.g. diabetes) and the lack of apparent benefit (to the producing individual or group) in releasing well annotated data into the public domain. This problem is very complex to tackle in an atomistic way; there needs to be a coordination of policy making, IT support and culture change to achieve openness. However, we would argue that Information Technology is currently lagging furthest of the three and has the most potential to support such a change.

A shift from data generation to knowledge creation

In the model of expansive learning, the community is in the centre-point of knowledge-creation. It is not sufficient for individual players to interact; but to co-create shared artefacts of knowledge. When we speak of such "trialogical" [26] mode of learning, we refer to combining the following three elements: (a) individual competencies, (b) communal participation and (c) co-creation. The danger of focusing on the first two without the latter is that knowledge may become reductionist, not expansive. When researchers tackle scientific questions in isolation or rely on a community to help answer them, inquiry is reduced to an individual mental process or a social process of participation. In "trialogical" learning, individual initiative serves the communal effort to create something new, and the social environment feeds individual initiative and cognitive growth. Thus, communal knowledge becomes materialized as common objects of activity are developed. An application of the above principles has powered the creation of the Knowledge Practices Environment (KPE) platform [27], used successfully in numerous collaborative learning projects [28,29,30].

Given time and resource restrictions few software providers can afford to develop from scratch. Thankfully, the open-source movement has now reached the mainstream and with immense success. If we are to expand on current knowledge using funds

in an efficient way, our utilisation of software cannot be any different. Recently available software toolkits such as Jango, jQuery, Google tools such as Google Docs or Google Refine and collections of widgets such as Bootstrap or BioPortal [31], to name a few, make it easy to develop functional software with little overhead. Even better yet, existing biomedical information management suites such as SIMBioMS [32], ISA tools [23], OBiBa (obiba.org) or BioMart [33] implement much of the functionality that will be discussed later. The main focus of software in the biomedical segment should therefore be integration, adherence to standards and interoperability, rather than lengthy generalised solutions, which risk being rendered obsolete as high-throughput technologies progress. If facilitating knowledge creation in highly complex environments is the goal, software groups should think more about designing for integration and communication rather than for solutions to problems.

DESIGN IS FUNCTION

In this section we wish to introduce recent advances in the design field, relevant to collaborative biomedical communities, such as open peer-to-peer design [34]. This design methodology is a promising community-based organizational form, building short and long collaborative networks with high probabilities of achieving success in society. Popular examples of similar efforts include Linux, Wikipedia, YouTube and other Web 2.0 communities. They represent, maybe, the only participation-based organizational forms with high scalability: the more the participants, the faster they achieve success. High participation however, means high complexity and to understand such complexity means to design in and for complexity itself. The Linux community arguably succeeded in this, because it faced the challenge of designing an operating system without reducing it, but by leveraging its own intrinsic complexity. The complexity of the project thus reflects the complexity of the community, and both strengthen each other.

The open peer-to-peer design process is a co-design process, where designer and participants collaborate in a wider design community (a collective intelligence). Designing software in such a context is an *enabling* act; not an act of delivering a solution. The chances of coming up with a better design are higher in such a scenario; and the better the design the more use the community extracts out of it. Thus design and function become synonymous terms.

COLLABORATIVE SOFTWARE IN BIOMEDICAL SCIENCES

We envision the next generation of information management software in the life sciences to be inspired by the following ideas:

Enabling complexity. Scientific communities should be given tools to self-organise and harness collective intelligence. These tools should not only support the exchange of data and information, but also enable the co-creation of knowledge, for example by means of group document editing, group discussion or distributed annotation.

A highly ethical exchange. While it is beneficial for data, information and knowledge to flow freely within a scientific community, there need to be means for acknowledging authorship of these scientific artefacts in a trackable way.

Standards, integration and networking. Each knowledge artefact needs to be organised into an appropriate context. Adherence to standards, such as minimum annotation requirements or exchange formats and integration with existing resources, such as through the use of URIs should be the bare minimum. Developing a network of partners can also enhance this process, as expertise and software components can flow in and out of the network for the benefit of integrating the most relevant software modules.

Commitment to openness. Open-source software and online collaborative communities are an unprecedented example of how Information Technology can harness collective intelligence. The next generation of information management software in biomedical studies should promote this idea in this field as well, trying to alleviate the shortcomings previously outlined.

We now shift our focus on how these requirements can be put into practice through the design of an IT infrastructure to support collaborate biomedical projects.

Software design in biomedical studies

Since the demand for distributed data management software grows constantly -both in terms of number of project and number of knowledge domains- it has become possible to present a consensus workflow. It includes researchers, data managers, information administrators (e.g. handling confidentiality information) and system administrators. If a discussion of such a generic workflow can be taken beyond the specifics of a particular knowledge domain, it can pave the way to creating a more sustainable communication infrastructure for the research community.

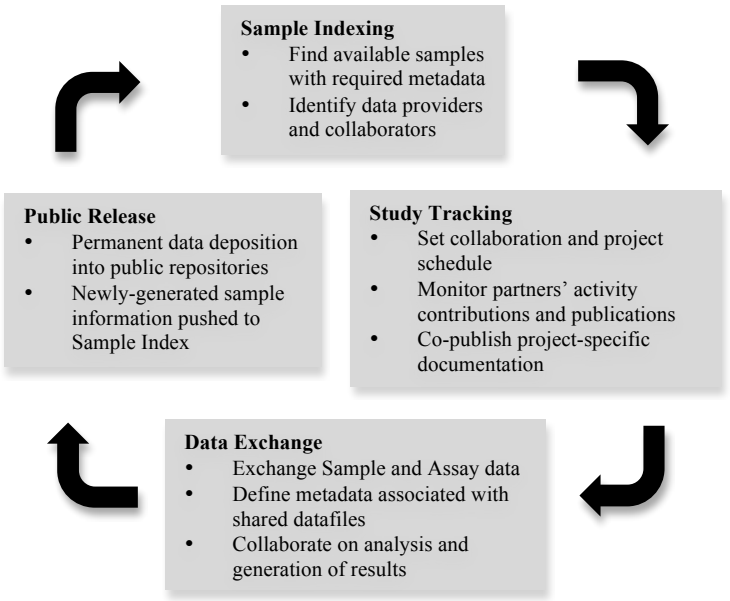

Figure 1. A set of common data management use-cases in biomedical studies, modelled into four distinct modules: Sample Indexing, Study Tracking, Data Exchange and Public Release

3.1.1. Sample Indexing. Research coordinators, looking for partners that can provide samples with high quality annotation that fulfils the needs of a study, start by querying a publicly accessible sample database. After selecting the list of required parameters (specific phenotypes, availability of

genotype data, etc) from the query interface, researchers are presented with a report that will show the number of available samples that satisfy the list of requirements from all registered data providers. Using the same system, the research coordinator is able to contact the data providers in question to initiate a collaboration and gain access to the sample data.

3.1.2. Study Tracking. On definition of study collaboration, the research coordinator makes use of a study-tracking module to add the list of participants to the project and to establish the schedule for the project. The study coordinator and the researchers are able to follow the progress of the project and monitor the activity of each of the partners. Each collaborator also uses this tool to publish and retrieve procedural information related to the study (protocols, publications, description of study, data access application, etc). The system will also allow study coordinators to add or remove partners and to set the access rights for each partner. Investigators may interact with various modules of the software depending on their role; *e.g.* biomaterial management, lab analysis, data analysis, *etc*. to access multiple services (databases, sftp, mailing lists, *etc.*).

3.1.3. Data Sharing. This module helps research coordinators to define their required data structure and system configuration through personal consultations, the study of data files and the capture of critical metadata descriptors. The product of this initial analysis is a customized installation of sample and/or assay databases with web forms and standard templates for data capture. Changes to web forms and vocabularies can be made by a researcher or an administrator using the graphical user interface. Configuration changes can be made quickly, making it possible to optimize the system in a timely manner. Once the system is configured, data upload/download and browsing is straightforward. Users can upload data manually using input web forms or they can use a customised template to batch upload sets of files of raw or processed data.

Additionally, analysis and visualisation pipelines can be connected to the data module as plug-ins. These plugins cater to needs for specific pre-processing, quality control and production pipelines, but also to individual research labs, as a means to publish in-house tools in a consistent and shareable manner. Results of these tools are fed back into the database and shared within the research community.

3.1.4. Public Release. Due to the requirement of many scientific journals of public access to research data, data sharing modules provide a solution to export the selected information directly to public repositories, such as the European Genome-phenome Archive, eliminating the need for the user to resubmit all the data to the final public repository manually. Ideally the use cycle of a biomedical study will close with the addition, by the research coordinator, of any newly generated sample data back into the Sample Index. This will increase the chances of reusing the sample data in new projects, improving the return of investment to public research funds by maximizing the number of publications that come out of a set of data.

CONCLUSIONS

Availability of Information Technology to enable it, pressure from public and industrial domains that have already adopted it, as well as increasing complexity, costs and demand for sustainability, are calling for open innovation in the life sciences. While there are tremendous benefits from harnessing the collective intelligence of communities to create knowledge from data collaboratively, there are disincentives in place to prevent open innovation from hitting the mainstream in life science research. Most notably, increased competition, fear of infringement of legal or ethical regulations, such as subject consent, or improper acknowledgement to the group that generated the research data. We believe IT can be in the forefront of the push for open biomedical science, by designing software pervaded by the following principles: enabling complexity, supporting ethical exchange of information, adhering to standards, integrating and networking whenever possible and being committed to openness by actively advocating it. We have provided a use-case information flow based on our experiences with data-intensive collaborative projects in the life sciences and hope to generate awareness and support in the biomedical software design community for enabling and integrating such beneficial collaborative research workflows.

ACKNOWLEDGEMENTS

We would like to thank Yulia Tammisto and Ugis Sarkans for fruitful discussions, advice and guidance. We also thank entire simbioms.org network for hard work and systematic gathering of usage data from 9 EU projects.

Conflict of interest statement: None declared.

Funding

This work has been funded by IP EC projects SIROCCO (LSHG-CT-2006-037900) and ENGAGE (grant agreement No: 201413).

REFERENCES

[1] T. Thorgeirsson, et al. **Sequence variant at CHRNB3-CHRNA6 and CYP2A6 affect smoking behaviour**. Nature Genetics (2010), 42(5):448 - 453

[2] M. Kolz, et al. **Meta-Analysis of 28,141 Individual Identifies Common Variant within Five New Loci That Influence Uric Acid Concentrations**. PLOS Genetics (2009), 5(6):e1000504.

[3] I. Prokopenko, et al. **Variants in MTNR1B influence fasting glucose levels**. Nature Genetics (2008), 41:77-81.

[4] A. Ripatti, et al.: **Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts**. Nature Genetics (2008), 41:47-55.

[5] A. Burgun, O. Bodenreider. **Accessing and Integrating Data and Knowledge for Biomedical Research in IMIA Yearbook 2008: Access to Health Information**, (2008) pp. 91-99.

[6] S. Oster. **CaGrid 1.0: an enterprise grid infrastructure for biomedical research**, JAMIA 15 (2008), pp. 138–149.

[7] P. McConnell, et al. **The cancer translational research informatics platform**. BMC Med Inform Decis Mak (2008) 24;8:60.

[8] C.F. Taylor, et al. **Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project**. Nature Biotechnology (2008), 26(8):889-896.

[9] B. Smith, et al. **The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration**. Nat. Biotechnol. (2007), 25:1251–1255.

[10] D.B. Keator, et al. **Derived Data Storage and Exchange Workflow for Large-Scale Neuroimaging Analyses on the BIRN Grid**. Front Neuroinformatics. (2009) 3:30. Epub 2009 Sep 7.

[11] I. Fortier, et al. **Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies**. Int. J. Epidemiol. (2010), 39:1383-1393.

[12] S. Brand. **The Media Lab: Inventing the Future at MIT**. New York: Viking/Penguin, 1987.

[13] N.R. Anderson, et al. **Issues in Biomedical Research Data Management and Analysis: Needs and Barriers**. J Am Med Inform Assoc. (2007), 14(4):478–488.

[14] H.W. Chesbrough (2003). **The era of open innovation**. MIT Sloan Management Review, 44 (3), 35–41

[15] **Gift System for Science**. Nature, Vol. 480, No. 7376. (7 December 2011), pp. 281-281.

[16] Molloy JC. **The open knowledge foundation: open data means better science**. PLoS Biol. 2011 Dec;9(12):e1001195. Epub 2011 Dec 6.

[17] A. Brazma, et al. **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data**. Nat Genet. 2001 Dec;29(4):365-71.

[18] N. Le Novère, et al. **Minimum information requested in the annotation of biochemical models (MIRIAM)**. Nat Biotechnol. 2005 Dec;23(12):1509-15.

[19] C.F. Taylor, et al. **The minimum information about a proteomics experiment (MIAPE)**. Nat Biotechnol. 2007 Aug;25(8):887-93. Review.

[20] A. Brazma, et al. **ArrayExpress--a public repository for microarray gene expression data at the EBI**. Nucleic Acids Res. 2003 Jan 1;31(1):68-71.

[21] M. Kapushesky, et al. **Gene expression atlas at the European bioinformatics institute**. Nucleic Acids Res. 2010 Jan;38(Database issue):D690-8. Epub 2009 Nov 11.

[22] D. Field, et al. **'Omics Data Sharing**. Science (2009), 326(5950):234-236.

[23] T.F. Rayner, et al. **A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB**. BMC Bioinformatics (2006), 7:489.

[24] P. Rocca-Serra, et al. **ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level**. Bioinformatics (2010), 26(18):2354-2356.

[25] M. Gostev, et al. **SAIL: A software system for sample and phenotype availability across biobanks and cohorts**. Bioinformatics (2011), 27(4):589-591.

[26] S. Paavola and K. Hakkarainen. **The Knowledge Creation Metaphor - An Emergent Epistemological Approach to Learning**. Science & Education (2005), 14(6):535-557.

[27] H. Markkanen, at al. **The Knowledge Practices Environment: a Virtual Environment for Collaborative Knowledge Creation and Work around Shared Artefacts**. In J. Luca & E. Weippl (Eds.), Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications (2008) pp.5035-5040. Chesapeake, VA: AACE.

[28] K. Hakkarainen. **Three generations of research on technology-enhanced learning**. British Journal of Educational Technology (2008). Volume 40, Issue 5, pp.879–888, September 2009

[29] A. Lund. **Assessment made visible: individual and collective practices**. Mind, Culture, and Activity (2008). 15:32-51.

[30] L. Lundvoll Nilsen and A. Moen. (2008). **Teleconsultation – collaborative work and opportunities for learning across organizational boundaries**. Journal of Telemedicine and Telecare, 14, 377-380.

[31] N.F. Noy, et al. **BioPortal: ontologies and integrated data resources at the click of a mouse**. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W170-3. Epub 2009 May 29.

[32] M. Krestyaninova, et al. **A system for Information Management in BioMedical Studies-SIMBioMS**. Bioinformatics (2009), 25(20):2768-2769.

[33] D. Smedley, et al. **BioMart – biological Queries made easy**. BMC genomics (2009), 10:22.

[34] M. Menichinelli. **openp2pdesign.org_1.1. Design for Complexity**. (2008), openp2pdesign.org.