DCC | Digital Curation Centre Case Studies and Interviews

**Integrative Biology
(http://www.integrativebiology.ox.ac.uk/)**

_____

Martin Donnelly, Victoria Boyd, and Jill Spellman
HATII, University of Glasgow,
Glasgow G12 8QQ
http://www.hatii.arts.gla.ac.uk

April 2008
Version 1.0

Digital Curation Centre Case Studies and Interviews: Integrative Biology

Catalogue Entry

Title	DCC Case Study – Integrative Biology
Creator	Martin Donnelly, Victoria Boyd, Jill Spellman (author)
Subject	Data curation; formats, processes and issues; system development; standards; legal factors; methodology, and problems overcome; human factors
Description	Computer simulation organ functions based on models offers the potential to increase researchers' understanding of the causes of medical conditions such as heart disease and to work towards developing new drugs and treatments to combat these illnesses. The main aim of the Integrative Biology (IB) project is to realise this potential by developing multi-scale models - spanning the range from genes to whole organs - and to provide data management features for its disparate users including the sharing of data in a secure infrastructure, and enabling the storage and re-use of simulation outputs.
Publisher	HATII, University of Glasgow; University of Edinburgh; UKOLN, University of Bath; Council for the Central Laboratory of the Research Councils.
Contributor	Joy Davidson and Andrew McHugh (editors)
Date	03 April 2008 (creation)
Type	Text
Format	Adobe Portable Document Format v.1.3 (Acrobat 4.x)
Resource Identifier	ISSN 1749-8767
Language	English
Rights	© HATII, University of Glasgow

Citation Guidelines
Donnelly M, Victoria Boyd, Jill Spellman (March 2008), "Integrative Biology", *Digital Curation Centre Case Studies and Interviews*, Joy Davidson and Andrew McHugh (editors), Retrieved <date>, from http://www.dcc.ac.uk/resource/case-studies/integrative_biology

Digital Curation Centre Case Studies and Interviews: Integrative Biology

About the DCC

The JISC-funded Digital Curation Centre (DCC) provides a focus on research into digital curation expertise and best practice for the storage, management and preservation of digital information to enable its use and re-use over time. The project represents a collaboration between the University of Edinburgh, the University of Glasgow through HATII, UKOLN at the University of Bath, and the Council of the Central Laboratory of the Research Councils (CCLRC). The DCC relies heavily on active participation and feedback from all stakeholder communities. For more information, please visit www.dcc.ac.uk. The DCC is not itself a data repository, nor does it attempt to impose policies and practices of one branch of scholarship upon another. Rather, based on insight from a vibrant research programme that addresses wider issues of data curation and long-term preservation, it will develop and offer programmes of outreach and practical services to assist those who face digital curation challenges. It also seeks to complement and contribute towards the efforts of related organisations, rather than duplicate services.

DCC Case Study and Interviews

Editors
Joy Davidson
Humanities Advanced Technology and Information Institute, University of Glasgow (UK)
Andrew McHugh
Humanities Advanced Technology and Information Institute, University of Glasgow (UK)

Peer Review Board
Michael Day
UKOLN, University of Bath (UK)
Guy McGarva
University of Edinburgh, (UK)
Mags McGinley
University of Edinburgh, (UK)
Maureen Pennock
UKOLN, University of Bath (UK)

Integrative Biology
(http://www.integrativebiology.ox.ac.uk/)

This Digital Curation Centre case study is the result of an email questionnaire completed by Sharon Lloyd and Daniel Hanlon of Integrative Biology, supplemented by subsequent email communications between the correspondents and Martin Donnelly and Jill Spellman of the DCC and HATII, University of Glasgow.

Table of Contents

Executive Summary

Computer simulation organ functions based on models offers the potential to increase researchers' understanding of the causes of medical conditions such as heart disease and to work towards developing new drugs and treatments to combat these illnesses. The main aim of the Integrative Biology (IB) project is to realise this potential by developing multi-scale models—spanning the range from genes to whole organs—and to provide data management features for its disparate users including the sharing of data in a secure infrastructure, and enabling the storage and re-use of simulation outputs.

'Many physiological models of the heart exist, but they are often restricted in scope by a lack of available computing power. In the case of tumours, a systematic modelling framework is yet to emerge. The biological modelling community is not like that of Particle Physics where the use of high performance computing is commonplace. It is not unheard of for a researcher's laptop to be the most powerful machine on which a simulation is ever run.'

Hanlon, D., L. Sastry and K. Kleese van Dam, "Integrative Biology at CCLRC"[1]

Introduction

The Integrative Biology (IB) Project could ultimately lead to improved understanding and treatment for heart disease and cancer—two of the UK's biggest killers, accounting for nearly 61% of all deaths.

One of the main aims of the IB Project is to exploit e-Science and Grid computing technology, harnessing the power of distributed computer and data resources to run ever more complex *in silico* simulations: multi-scale models from cell level to whole organs, including the assimilation of *in vitro* and clinical data to fine tune the models. The Integrative Biology Grid enables experimental and computational biologists to exploit the coordinated use of supercomputing resources to run simulations of these models, with spatial and temporal resolution not possible prior to the e-Science Programme.

A key driver is the e-Science vision of Grid technology being used by computational biologists to carry out collaborative simulated experiments within a distributed virtual laboratory, or 'collaboratory.'[2] Such a lab will integrate the best available resources for computation, data management, visualisation and analysis. In running complex multi-scale computer models of whole organs, the keys to progress are better accuracy and greater detail, coupled with enhanced computing power.

This project began in February 2004 and is international, multi-institutional and multi-disciplinary. The project partners comprise leading biomedical and computing experts from: University of Oxford, University of Nottingham, University of Birmingham, University of Sheffield, University of Auckland, University of Graz, University of Baltimore, University of St Louis, University of Calgary, Ultrecht University, Tulane University, University of California, San Diego and University of California, Los Angeles with the University of Leeds, University College London and the Science and Engineering Facilities Council (STFC, formerly known as the Central Laboratory of the Research Councils, or CCLRC). Sponsors include STFC, the Engineering and Physical Sciences Research Council (EPSRC), and IBM.[3]

The project has received £2.44M funding from the e-Science Programme, through EPSRC, to support ten post-doctoral research associates (PDRAs) and six PhD students. The University of Oxford has provided funding for a further four PhD students, plus system administration and secretarial support. The Universities of Nottingham and Birmingham each fund one additional PhD student. STFC provides the use of its Atlas DataStore (ADS) data archiving facility, a database server cluster at the Rutherford Appleton Laboratory, plus additional staff time.

[1] Hanlon, D., L. Sastry and K. Kleese van Dam, "Integrative Biology at CCLRC," *ERCIM News*, No 60, January 2005. URL: http://www.ercim.org/publication/Ercim_News/enw60/hanlon.html

[2] http://www.worldwidewords.org/weirdwords/ww-ins1.htm
[3] http://www.integrativebiology.ox.ac.uk/sponsors.html

The project enjoys extensive access to national supercomputing resources at HPCx.[4] In addition to the main project staff, the project also benefits from a large pool of expertise comprising co-investigators from the project partners. Existing links with the pharmaceutical and biotechnology industries are strengthened through an ongoing needs analysis study of evolving users through annual two-day workshops with the technical and scientific communities[5], as well as through annual colorectal cancer modelling workshops designed to bring modellers and experimentalists together.[6]

The IB project team consists of four main groups developing the:

- Modelling and simulation codes;
- Computational and data management framework for simulation and integration;
- Advanced visualisation techniques for complex datasets;
- Advanced coupling techniques and steering mechanisms for models.

Within these groups, cross-institutional teams are working on specific technical areas including:

- Heart modelling;
- Cancer modelling;
- Molecular and cellular modelling;
- Testing, tuning and running simulations;
- Job management in a distributed high performance computing environment;
- Data management;
- Computational steering;
- Workflow;
- Visualising data and user interfaces.

Models and Simulation

The project's long-term goal is to develop a theory of biology and biological function with a system capable of computational analysis that is sufficiently accurate to produce clinically useful results.[7] The components and processes in the chain which must be modelled and integrated by the IB systems range from genetic information to cells and tissue, and finally to the behaviour of whole organs.

IB aims to enable several key features for data management for its disparate users including:

- The provision of a secure infrastructure in which users can store valuable and sensitive information, with the ability to control access privileges to these resources.

- The ability to enable data sharing selectively through the concept of an 'experiment' with experiment PI controlling access rights (a feature of the next release of a Storage Resource Broker (SRB)).

- The ability to leverage prior work—that is already in a mature stage—by working with group experts in the UK, such as STFC.

- The ability for advanced, specialist server-side visualisation applications to access the data securely, using third party data

---

[4] The UK's national high-performance computing service, which is provided by EPCC at the University of Edinburgh and by CCLRC Daresbury Laboratory, and funded by the Office of Science and Technology through EPSRC's High End Computing Programme, http://www.hpcx.ac.uk/

[5] The annual two-day workshops are by invitation only and include representatives from other research institutions in addition to a consultant who is working extensively with the pharmaceutical industry in relation to drug testing. The consultant works closely with Denis Noble FRS and other eminent scientists on an Fp7 grant application, which is currently seeking funding to explore the use of technology and modelling for drug reintroduction.

[6] The colorectal cancer modelling workshops are by invitation only and discuss potential future work together.

[7] http://www.ercim.org/publication/Ercim_News/enw60/mac_randal.html

transfer techniques and SRB interface, to create visualisations that are then delivered to the users' desktops for data analysis.

- The ability for simulations to store outputs from simulation runs using third party data transfer techniques on SRB, including ADS using the SRB interface.

With these specific features in mind, the project consortium endeavoured to bring together groups with appropriate expertise. Data management and existing practices of the scientists were examined during the initial requirements exercise and, in many cases, archiving was either scant or non-existent. As a result a risk register has been created and maintained.

The project's aim is to develop multi-scale models of organs and tumours, and to run simulations of these models via the coordinated use of supercomputing and distributed data[8] resources. Grid computing allows simulations at a spatial and temporal resolution that was impossible prior to the e-Science Programme. By using the most powerful supercomputers available for research in the UK today, IB is gradually improving the understanding of these two complex systems, leading to improved control and treatment regimes. Essentially, IB has developed a prototype e-Infrastructure, an IB Grid that facilitates the collaborative development and secure exchange of simulation codes created by research teams from around the world.

The major stakeholders and beneficiaries in the IB Project are the participants and their host organisations, EPSRC and the UK e-Science Programme, the UK biomedical community and those who will benefit from the advances that the

project will deliver in the understanding of heart disease and cancer.

System Development
Three of the PDRAs (STFC) are dedicated to core technology development, with an additional RA looking at advanced visualisation techniques for heart modelling (Leeds) and another RA exploring the coupling of models (UCL). A key challenge is scheduling access to the appropriate combinations of distributed resources that are needed to run sophisticated organ simulations, in a transparent way.

An overall architecture for the project's software infrastructure has been designed as the basis of current development work. A simple overview of the project's initial architecture and the different levels of software infrastructure is given in Illustration 1, below.

Illustration 2 shows the system architecture including individual functional components. This architecture is a refinement that has evolved over the past three years as users' requirements have changed. Re-evaluation of needs has enabled implementation of the timely visualisation of results, with heart modelling visualisation-specific packages integrated into the core IB Infrastructure in a scalable way.
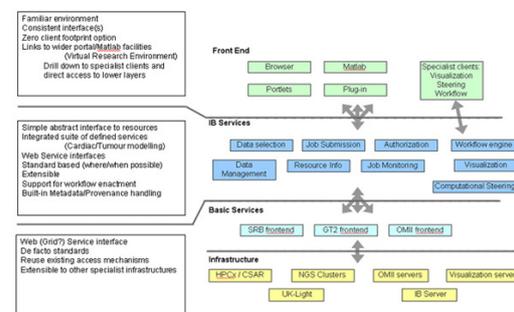

Illustration 1: Architecture overview

---

[8] Many of the input files to the simulations are large files. These are also stored in SRB for access by various simulation runs.
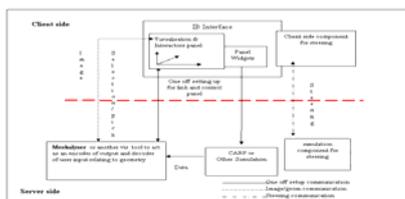
Illustration 2: Detailed system architecture diagram

The project architecture comprises five main components:

- Infrastructure components. These provide user-accessible services such as the preparation, submission and monitoring of jobs, computational steering, managed data access, and control of data visualisation;

- A simulation subsystem, providing 'solvers' for user-supplied model codes;

- A data management subsystem. This stores the user's data files and associated metadata, and provides facilities for data retrieval and curation/annotation;

- A visualisation subsystem, offering a range of visualisation techniques for examining simulation and experimental results. This can run in tandem with remote colleagues, allowing and facilitating real-time collaborative work;

- The user interface. This runs within a standard Web browser, as well as a stand alone IB Interface (IBI) for all the generic services listed above and couples that with the control interfaces of the visualisation toolkits used on the server side. The latter toolkits include specialised desktop applications such as Meshalyser, Matlab etc.[9] IBI provides a SRB interface to SRB and ADS while the user interfaces use SRB clientside library to access SRB and ADS.

Simulations are currently being run on a range of machine types, from standard local workstations, to the Grid's commodity workstation clusters, to the UK's very high performance supercomputing facility at HPCx at the University of Edinburgh.[10] A SRB is used to manage the wide variety of datasets generated at multiple locations in the project.[11] Information visualisation can either be carried out either locally or using Grid-based resources, depending on complexity.

The primary user tasks supported by the software infrastructure are:

1. Incorporating computable models of biological systems into executable simulation codes. These can be installed on a range of computer systems;
2. Specifying the input parameters of a simulation problem;
3. Selecting the resources to be used and running the simulation;
4. Monitoring and controlling the simulation during its execution;
5. Managing and curating both input and output datasets securely;
6. Creating appropriate metadata information to enable future reference and re-use;
7. Analysing simulation results and experimental data, assisted by data visualisation;
8. Storing intermediate and end results in images for publications, including animations, as data files, using structured data management on the SRB and ADS;
9. Metadata about simulation runs are stored on the National Grid Service (NGS) Oracle database as persistent objects for tracking and browsing experiments;
10. Assisting in collaborative, distributed work regardless of location.[12]

---

[9] http://www.mathworks.com/

[10] http://www.hpcx.ac.uk/
[11] http://www.sdsc.edu/srb/index.php/Main_Page
[12] http://www.integrativebiology.ox.ac.uk/softarch.html

The development team is guided by three key considerations:

- The use of established standards frameworks, which they help to develop where appropriate;

- Ensuring that the software they develop is scalable, to address the need for increasing spatial and temporal resolution;

- The framework being built must be secure enough to protect the integrity and confidentiality of the project's assets.

The developers have created prototype systems, with research scientists closely involved in order to define user requirements and to provide feedback on the trial applications. The project intends to release a new prototype every six months, built atop the rapidly developing Grid middleware.[13] In order to meet its goals, IB is creating a stable middleware environment, compatible both with existing codes and with the Grid philosophy, and synthesising existing technology components from various e-Science projects. This approach reduces duplication of effort by promoting the re-use of code and expertise.

Technical Challenges

IB is a second-generation e-Science project, building on developments from first generation projects. In turn, IB is itself advancing the e-Science agenda on several fronts. Challenges for the project in this area include:

1. Providing transparent, scheduled access to the combinations of distributed high performance computers and database resources that are appropriate for running multi-scale organ simulations;

2. Exploiting these resources efficiently by applying computational steering, workflow, visualisation, and other techniques developed in previous e-Science projects;

3. Enabling globally-distributed researchers to control, analyse and visualise simulation results collaboratively, thereby progressing the project's scientific agenda;

4. Maintaining a secure environment for the resources used and information generated by the project, without inhibiting scientific collaboration.[14]

A major factor in the choice of technologies for IB data management has been the provision of a Grid Infrastructure by the NGS. To a large degree the grid technologies deployed by the NGS have already been integrated though project specific requirements to provide added functionality. The choice of a Grid Service means that the IB team are deploying onto a software infrastructure, which has emerged as the preferred choice of the NGS Grid Operations Centre, and based on the findings of the Grid Engineering Task Force, which was funded to analyse and comment on Grid Computing technologies.

Data Curation

Current IB users are predominantly data creators, running numerical codes to model physiology. As the IB

---

[13] The first prototype release (June 2005) includes: (a) job submission of model runs to NGS via a portal based on the CCLRC portal; (b) security based on NGS certificates using myProxy; (c) data management using SRB to hold simulation input and output files; (d) visualisation provided by tools for specific modelling requirements; (e) code management using CVS. The prototype is accompanied by user documentation, and training is provided as required. Demonstration screenshots and explanations are available online at http://www.integrativebiology.ox.ac.uk/demos.html

[14] Adapted from http://www.integrativebiology.ox.ac.uk/challenge.html

infrastructure develops, it is envisaged that new numerical physiologists will become re-users, accessing the stored results of previous calculations for further analysis.

Users hold their data in the SRB provided by the NGS.[15] In this developmental stage of IB, users manage the data structure. The location and content of data is self-documented and the resulting documentation self-managed. The greatest proportion of the data is raw output from numerical codes. Longer-term preservation/curation of data is handled by the ADS facility at the Rutherford Appleton Laboratory.[16] A connection has been established between SRB (the main, spinning disk data repository) and the ADS for archival purposes. A strategy to determine when particular files are archived is being developed within the context of a more automated data management infrastructure, where IB tools disassociate (or 'abstract') the end user from details regarding the physical whereabouts of individual data.

Standards and Legal factors

Whilst the principal development collaborators and their contributing partners will sign a collaboration agreement which covers confidentiality, the main user groups trust the project to implement a solution that protects their assets. Much of the work involves voluntary sharing of data and software, and discussions on open source licensing are actively encouraged and supported through the JISC-funded OSS Watch group. However, providing access to the raw data is often less sensitive than access to lab-books and the series of actions/methods which may reveal a researcher's scientific agenda.

The project employs the Grid Security Infrastructure throughout.[17] In conjunction with the UK e-Science Certification Authority's X509 certificates, the IB environment can therefore claim strong user authentication and encryption capabilities.

Methodology, and Problems overcome

The overall methodology for deploying solutions in IB is an iterative process informed by requirements elicitation and review processes. The project aims to leverage output from many other projects and to provide incremental delivery across workpackages. Constraints can sometimes be imposed by the development or implementation timelines of the partnering organisations. That said, IB has released incrementally, with a major showcase to users and new potential users given in March 2007. Since then, closedown activity has focused on delivering a refined system called IBI (IB Infrastructure), taking into account the need to produce geometry for visualisation on a specialist visualisation cluster on the Grid (data sets were becoming too large to download to the desktop to visualise) and the need to incorporate the users' favourite visualisation tools. This approach was chosen to ensure 'buy-in' from the user community who require constant reassurance that technology development keeps abreast of the needs of the scientific community.

By focusing initially on deploying an existing solution, a solid base has been established on which the rest of the project can grow. Data management is the crux of many other aspects of the work being undertaken, underpinning the process of running simulations and of visualising results. It is often the case, however, when

---

[15] The total usage is currently 331GB.
[16] IB also has extensive access to national HPC resources (e.g. HPCx).

[17] http://www.globus.org/security/overview.html

utilising generic resources, that implementation is slower than if one had total control over the equipment and infrastructure. But by developing in this way, the project is feeding into the usability and effectiveness of generic solutions for a wider audience.

Human Factors

The STFC e-Science Centre provides a large proportion of the personnel developing the software infrastructure for the IB Project. This has greatly simplified the process of deployment, as many of the technologies would already be familiar to the developers. Use of the NGS infrastructure and the close working relationships of IB developers and NGS support engineers should make further expertise available as it is needed in the future.

Evaluation and Future Work

IB has ongoing open discussions with the user community as software develops and new features are released, and is therefore able to accept feedback and act on it as required. Initial releases were distributed to a limited audience who reviewed the implementation and provided feedback for refinement before major releases to the rest of the IB user community in 2007.

Researchers have already been able to model highly complicated biological systems. In time they will be able to "grow" virtual tumours through the crucial stages of early development, and in doing so gain a greater understanding of the complex biochemical and physiological processes that take place during tumour growth.

Long-term archiving is an issue already in progress through some of the funding councils. It is widely recognised that these principles need to be embedded into funded science in order to ensure that the results and

data generated can be retrieved post-project, indeed even after the data generators have moved on. The key point is that having the opportunity for data re-use is as important as the initial generation of the data.

As the project continues, the team will review future use of SRB across continents by looking into ways of federating catalogues, thereby reducing the need for mass implementation of global metadata. In the longer term, it is hoped that IB project research will lead to an improved understanding of biological systems. The project foresees "a future where new drugs will be discovered and tested using computer models such as those which we are developing."[18]

---

[18] http://www.integrativebiology.ox.ac.uk/

Digital Curation Centre Case Studies and Interviews: Integrative Biology

ANNEX — Project work programme

The IB work programme is organised into nine work packages:[19]

WP1 Executable management
A personalised executable registry from which users can select executables for use in a simulation or workflow. Provides access to personal executables, as well as executables available in experiments and public registries. The project is also providing access to software development tools supporting modelling and code creation, particularly tools or instructions to support deployment. Matlab and CMISS are being supported and integrated into the project infrastructure, and the use of mark-up language CellML and other component repositories is being facilitated.[20]

WP2 Workflow management
This mirrors and complements WP1, providing facilities to add, edit, delete, publish, search, review and select workflows, including the relevant metadata. Integration with the *myGrid* workflow designer and workflow engine offers facilities to create and execute these workflows.[21]

WP3 Job submission
WP3 enables users to set up input parameters for an executable or workflow run, and to submit jobs to project-accessible resources via a dedicated, personalisable portal. In time this will support specification of routing for input/output data, computational steering inputs and outputs, check pointing and failure handling. Where possible, appropriate metadata/provenance data will be created automatically. Users will be able to check and steer the status of a job (or workflow) during execution, either individually or in collaboration with other users.

WP4 Data management
This consists of services which allow users to manage their data space in the temporary, medium and long term, with automated backup and curation for specified data types, accompanied by the appropriate provenance metadata. A naming scheme/namespace is also being developed. This work package will enable users to 'publish' data for use by other scientists, as well helping them to search, review and select datasets according to their access rights. Likewise, users will be able to review those who have accessed their data, and to set up access rights to their datasets to ensure their appropriate use.

WP5 Visualisation
Users can access a variety of visualisation methods and tools, with associated metadata used to inform tool choice. Visualisation of results will be possible by one or more (possibly remote) users, both in real time and as the simulation is executing, if computational steering to monitor progress is embedded. Users can also create movies from time-slice data, and can control the viewing of simulation results either through rotation or movement in the data space or by varying display parameters. Remote users' viewpoints may be fixed (where everyone sees the same) or steerable (where each user can change his or her view independently of the others).

WP6 Computational steering
This provides tools for steering a job or workflow, enabling users to alter the model input parameters. This is achieved through real-time visualisation of results, and

---

[19] Condensed from http://www.integrativebiology.ox.ac.uk/workprog.html
[20] http://www.bioeng.auckland.ac.nz/cmiss/cmiss.php; http://www.cellml.org/
[21] http://www.integrativebiology.ox.ac.uk/insilico-project.html

subsequent parameter changes based either on real-time user input or preset parameter boundaries. Users will be able to spawn jobs while steering a simulation, allowing them to view the results of multiple simulations across a range of parameter values.

WP7 Collaboration and WP8 user interface

The IB services have the standard Web Services interface to the client side to accommodate the issues of firewall. Hence these services can be and are embedded into a variety of desktop toolkits. One such toolkit is the IB Virtual Research Environment (IBVRE) project[22], funded under the Joint Information Systems Committee's Virtual Research Environments programme. There is also a standalone C++ interface that can be embedded into other toolkits.

WP9 Security

Providing adequate data and code security is a high priority. Integrity and confidentiality must be maintained while providing managed access to data and codes. This raises issues of tolerance to component failure, requiring active mechanisms to ensure that key data can be protected against modification. The unit of security in the project is the *in silico* experiment comprising a set of users, resources, data and computations that form a coherent collection. The integrity of each experiment must be maintained, and protected against external modification. Within the experiment the different components need to be allowed to interact without the risk of compromising each other. To these ends, every experiment will have a designated principal investigator, responsible for defining its security requirements.

---

[22] http://www.vre.ox.ac.uk/ibvre/