

# VRE for the Integrative Biology Research Consortium (IB VRE)

**Matthew Dovey, Matthew Mascord**

Research Technologies Service, Oxford University Computing Services

David Gavaghan, Sharon Lloyd, Andrew Simpson, Geoff Williams

Oxford University Computing Laboratory

Damian Mac Randal

CCLRC

## Abstract

The IB VRE project is a JISC funded project developing a large-scale Virtual Research Environment demonstrator to investigate the use of existing collaboration frameworks to support the entire research process of a large-scale, international research consortium, namely that of the Integrative Biology (IB) project. IB is a second-round EPSRC e-Science Pilot project developing a Grid infrastructure to support post-genomic research in integrative biology, demanding a comprehensive range of research-support and grid-enabled tools and services. The IB community seek to further both their performance through collaboration addressing heart modelling and cancer modelling. IB VRE aims to provide a single, integrated environment supporting the entire research process from experimental and simulated data generation, acquisition, analysis and curation, through access to HPC and experimental resources, to project management, administration, and learning and teaching support tools. IB VRE is based on a WSRP and JSR168 compliant portal container allowing tool reuse (e.g. those provided by OGCE 2.0) and has a layered architecture. A key component is the support provided for the “research process” – including the “work” aspect of the process as determined by workflows, and the “social” aspects, by using existing tools for collaboration and project management. This paper outlines the motivations, usage scenarios and technical architecture proposed for this environment.

## 1. Introduction

The IB VRE project is a JISC funded project which is developing a large-scale Virtual Research Environment (VRE) demonstrator investigating the use of existing collaboration frameworks to support the entire research process of a large-scale, international research consortium [1]. In this case the consortium is the Integrative Biology (IB) e-Science Pilot Project [2], which requires a rich and complex environment, demanding a very comprehensive range of research-support and grid-enabled tools and services[3]. Hence it provides an ideal test-bed for evaluating and developing existing and planned VRE middleware. The IB project is developing integrated multi-scale whole organ models for some of the most complex biological systems in the clinical and life sciences; using these models to study the development cycle of cardiac disease and cancer tumours; bringing together clinical and laboratory data from many

sources to evaluate and improve the accuracy of the models; understanding the fundamental causes of these life-threatening conditions and how to reduce their likelihood of occurrence and identifying opportunities for intervention at the molecular and cellular level using customised drugs and novel treatment regimes.

## 2. Integrative Biology Project

The Integrative Biology (IB) project is a second-round EPSRC e-Science Pilot project which commenced in February 2004. The primary aim of the IB project is the development of the Grid Infrastructure required to support post-genomic research in integrative biology. Integrative or systems approaches to research in biology are evolving very rapidly, driven by a pressing need to understand how the components that make up a biological system interact across multiple spatial and temporal scales to allow biological function to emerge, and to express that understanding in quantitative

terms. It is now widely accepted that this can only be achieved by an iterative interplay between experimental data (both in vitro and in vivo), mathematical modelling, and computer simulation. The IB project is building the IT infrastructure to support this cycle: from experimental data, through the model-building process and HPC-enabled simulation, to experimental and simulation data capture, storage and analysis, and on to model validation and the subsequent design of new wet lab and in-silico experiments. This infrastructure is being built using a service-oriented Grid architecture, and is utilising many of the middleware developments within the UK e-Science Programme and in the wider Grid community, and within the JISC Middleware programmes.

### 3. IB-VRE

Initial research within the IB project has focused on developing a set of early demonstrators to evaluate the utility of existing tools and services developed by first-round e-Science projects to meet the aims of this diverse research community, and to act as a basis for the requirements gathering process [4]. A range of user interfaces to these tools are utilised, with the longer-term intention of developing a suitable Grid portal to the underlying infrastructure, leveraging extensive experience at CCLRC. The IB architecture provides an ideal test-bed for development of a comprehensive VRE to support a very large-scale, complex research project undertaken by a geographically distributed and interdisciplinary community. The JISC funded IB VRE project therefore aims to provide to the IB community a single, integrated environment to support the entire research process from experimental and simulated data generation, acquisition, analysis and curation, through access to IT, HPC and experimental resources, to project management, administration, and learning and teaching support tools.

IB brings together communities of scientists and technologists from across the globe who seek collaboration to further both their individual and group performance through collaboration. Collaboration often develops over time through the process of building trust and common understanding. Ackerman states [[5] p. 13,] that “interactivity is a key to learning” and “An increasing number of software designers, cognitive scientists and educators have come to the view that experience is actively constructed and reconstructed

through direct interaction with the world, and that, indeed, knowledge is experience”. Computer Supported Collaborative Work (CSCW) is a leading forum for presenting and discussing research and development achievements in the design, introduction and use of technologies that affect groups, organizations, communities, and societies. CSCW as a discipline is the study of how people work together using computer technology. It has been long recognised that whilst collaboration is a social activity, the technology to support this activity is crucial in ensuring its effectiveness. With the emergence of new CSCW collaboration tools [6], global collaboration is now feasible.

## 4. IB-VRE Usage Scenarios

This section details specific aspects of collaboration within the Integrative Biology project identified through our initial requirements capture exercise. These use cases will be refined over the life of the project through prototype evaluation with the disparate users.

IB is addressing two areas of medical modelling, heart modelling and cancer modelling. Collaboration between heart modellers has typically been through developed relationships between institutes e.g. Oxford's Denis Noble and Auckland's Peter Hunter. These collaborations have included periods of joint research, staff secondment and submission of joint papers over many years. Each has their own specific areas of expertise and interest and both understand that a combined programme of research leveraging the skills and interests of both groups would enable key research questions to be answered faster. Other key collaborations in the area of heart modelling have emerged over the life of the Integrative Biology through bringing these researchers together to discuss potential ‘experiments’ and publications. An example of such a relationship is that between a researcher in Oxford with her previous employer in New Orleans, as well as complimentary collaborative research between the Oxford researcher and colleagues in Graz, Austria and colleagues in Sheffield, UK. The following use cases cover the life cycle of such a relationship from its initial establishment through to the publication of results.

### 4.1 Identify papers and expertise in a selected area of research.

Often these areas of research are as a result of a gap analysis – ‘I need to understand ‘x’ to be

able to complete this experiment'. A simple scenario here would be a search on publication databases or Google. A more advanced scenario would be notification of new papers and publications based on a user signing up to an area of interest.

#### **4.2 Identify potential sources of funding for collaborative research.**

A simple scenario here would be a user signs up to an area of interest and receives notification either by email of an 'inbox' within their virtual research environment of possible funding opportunities matching their profile. An example of this would be the notification service from [www.rdinfo.org.uk](http://www.rdinfo.org.uk).

#### **4.3 Real time textual communication.**

A typical scenario would involve identifying person or persons to invoke a real time discussion with, inviting them to participate and proceeding to 'chat'. This scenario would be extended to include the sharing of visual collaboration through either collaborative results visualisation (use case 5) or real time video communication (use case 4)

#### **4.4 Real time video communication**

Use of personal access grid have proved to be of huge potential benefit to clinical researchers where mobility is problematic. On the another e-science project, the NeuroGrid project, these personal access grids enable users to invoke interaction which is more personal than telephone or conference call. The ability to show presentations at the same time is also a benefit. Both user cases 3 and 4 would normally be augmented by in person meetings as well as virtual meeting within the virtual research space.

#### **4.5 Real time visualisation and collaborative steering**

Key to the collaborative element of running a joint in-silico experiment will be the ability for several researchers to visualise, manipulate and potentially steer a model as it is running. This would require extensive technology support to enable the users to see large datasets in visual form, manage the control element of e.g. moving the image, changing parameters and steering the application. This activity would typically be coupled with real time textual or video communication.

#### **4.6 Manage workflows**

Reuse of processes will also form an essential part of the working practices of our scientists. Workflow will be an essential capability to enable users to record, replay and share common procedures. Scenarios may include activities to retrieve data, submit jobs to run on compute resources with specific parameters or conditions, or perform more administrative tasks within the research environment. Sharing of these workflows will enable colleagues to benefit from previous experience and best practise through the publication of such process flows.

#### **4.7 Managing Publications**

Key to any researchers career is the process of publishing the results of their research. Frequently publications will be collaborative with the development of an abstract or paper requiring multiple inputs to complete. Clearly document management and versioning will be vitally important to ensure that this process is effectively managed.

The other aspect of IB-VRE is cancer modelling. Computer modelling of tumour formation and growth is less advanced than that for the modelling of the heart. Collaboration has tended to be more in terms of carving out specific aspects of research for a group to undertake to create a niche for that department. Key to the success of this activity is the identification of existing publications through use case 4.1 as identified for the heart modelling activity. Notification here is considered vitally important. Similarly use case 4.2 would apply to the cancer modelling domain to enable the relevant groups to identify potential areas of funding. The cancer modelling community will learn from the experiences of the heart modelling community and will ultimately benefit from the additional capability highlighted above.

Collaborative research within the Integrative Biology project is not constrained to the scientists, in fact with disparate technology resources across the project and development activity also involving technologists who support the scientists from partner organisations outside of the UK, the VRE environment will also benefit these teams. Key uses will include collaborative planning, publication management, source code management and testing, as well as diary and calendar management.

The use cases above are a selection of those that are envisaged to form part of such a virtual research environment for the Integrative Biology scientists and technologists and these will be supplemented by the integration of the key services for the project including job submission and management, data management and visualisation as well as additional features to support the collaborative management of diaries and calendars, the concept of shared project knowledge including planning and administration. We will aim to learn from earlier developments in the area of virtual learning and expect elements of this existing capability to benefit both the scientists and the technologists on the project. The ability to capture and communicate knowledge effectively, as typified by the MIAKT project, where meetings were captured and annotated for future review, are essential to ensuring that key information is not lost. Core to the success of such an implementation will be developing tools and services that are usable and configurable by a diverse and global community, enabling the scientist to immerse themselves within this environment.

A key, and innovative, part of this project is the support provided for the “research process” both in terms of the “work” aspect and the “social” Aspect. In regards to the former, one of the central facilities of a VRE must be the ability to capture and reuse processes. In an IB “in-silico experiment”, the process takes the form of workflows, and, working with the myGrid project, we are developing tools to allow this workflow information, including provenance, to be saved, and that allow this information to be annotated, categorized and retrieved for reuse. In addition we are also capturing outcome information – was the workflow successful or not? This body of information is important because the knowledge of the experimental process is itself used to interpret and validate the results obtained. However, by capturing the workflow information we can also allow new users, or users with new requirements, to build upon a knowledge base of previous workflows. Coupled with an intelligent scoring or recommendation system based on machine learning techniques, these captured workflows

can therefore be used to represent best practice, and used either as the basis for future experiments (perhaps undertaken by less expert users), or as part of a collaborative or learning environment where colleagues or new researchers can learn by observation and re-enactment. This “work” aspect of the “research process” will be augmented by existing tools for managing the “social” aspects as described above (project management, publications, teaching etc.).

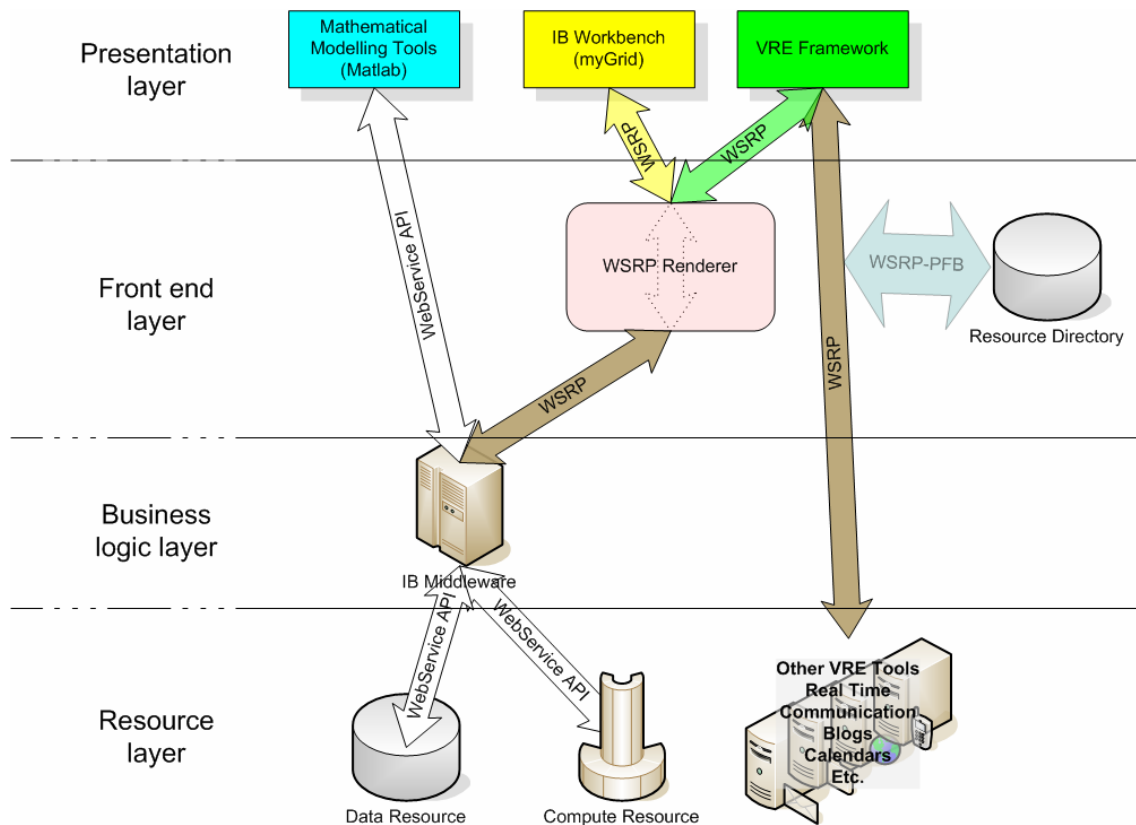
## 5. IB-VRE Architecture

The VRE is based on a WSRP [7] and JSR-168 [8] compliant portal container to ensure that the tools can be deployed within any portlet-based VRE environment. This will also allow relevant existing WSRP portlets and JSR168 portlets to be used within the IB VRE as well as portlets written specifically for IB. In particular we will be using and evaluating the JSR 168 Grid portlets provided by OGCE Release 2 [9]. We will also be evaluating the use of such portlets within other environments such as Matlab and myGrid.

In extending the service-oriented IB infrastructure, the IB VRE adopts a layered or tiered architectural framework: presentation; front end; business logic; and resources. The essential layers are:

### 5.1 Presentation layer

The user client must be lightweight, ideally just a web browser. However early experience within the IB project suggests that some of the more sophisticated user interface capabilities, in particular interactive visualization of complex result sets, will demand additional client-side assistance. In addition, personalisation and scaling requirements may also require client-side assistance. This assistance must not be mandatory in that users should be able to access and exploit the core VRE functionality with only a web browser. Behind the user client will be a portal server running on the IB infrastructure and providing access via coordinated portlets to the underlying facilities/services.



## 5.2 Front end layer

IB portlets, also running on the IB infrastructure, provide access to the core IB functionality: Resource Location, Job Composition, Job Submission, Computational Steering, Data Management (inputs and results), and Visualization. In addition, a Workflow portlet allows the user to merge and manage this functionality in terms of coherent in-silico experiments. (One of the key highlights of the IB system is capture of the workflow underlying the experiment, together with other automatically gathered metadata to provide full provenance information attached to the experimental results). To this set will be added portlets giving access to the extended range of domain-specific functionality.

## 5.3 Business logic

Behind the portlets are the underlying services. These can be internally layered in three strata, the coordination/orchestration layer, the functional services layer, and the base service layer. The latter services provide direct access to the computational and data resources, and will generally be hosted remotely. The middle layer aggregates and presents these as functional services relevant to and understandable by the user. Most of the work of the system occurs in

this layer and the user interacts directly with these services via their portlets and the portal. The upper layer helps the user manage these functional services, i.e. it supports the “experimental process”. It is in this layer that the major work of creating the VRE lies, moving the process support up from managing jobs to managing the user’s research environment. Another important component of this layer will also handle the management of provenance information. Within the IB project, the provenance of scientific results is clearly a major topic, and significant resources are being directed at automatically capturing as much relevant information about results as possible, and providing users with the opportunity to attach scientifically meaningful metadata and provenance data in addition. In the wider context of a VRE, additional facilities will be required, for example to relate results to the scientific papers which they support, or to help plan additional experiments to fill in gaps in the collected results. This will most likely be layered on top of the existing IB provenance/metadata infrastructure, with the VRE portal providing access and management facilities

## 5.4 Resource layer

The underlying resources in IB are computational, modelling and data storage

systems. Through the VRE, many other resources are made available. In particular, access to published and unpublished information (such as e-journals, e-print archives and raw data archives), project management tools (including project wikis and code repositories), Personal Information Management tools (diaries, contact lists, task management, etc), conference and messaging management, and teaching learning tools will be supported. In addition the VRE incorporates existing tools for managing and tracking the development process itself including project management tools, bug tracking tools and enhancement request tools; i.e. that as it the VRE is developed, the VRE itself will be used to maintain and develop the VRE.

## 6. Security

In an increasingly distributed but interconnected world, security is rapidly becoming a key requirement for all IT systems. Security is both a property of the design of the whole system and a consequence of the operational constraints built into each component. Portals, by their very nature, pose more than the normal share of security problems. Facing both a diverse set of users on one side and an equally diverse set of resource providers on the other, they have an added responsibility as a trusted intermediary and gatekeeper.

The fundamental security unit in the Integrative Biology project is the “in-silico experiment”, i.e., a coherent set of users, resources, data and computations focussed on achieving a specific goal or goals. The IB VRE will also adopt this approach, extending the notion of experiment beyond the science domain to address management and collaboration aspects of research. Individual experiments have to be isolated, or at least protected from interference from the outside, and individual components have to be allowed to interact within one or more experiments without compromising each other or the Experiments to which they are contributing.

As the natural gateway to the experiment’s components, the portal will play a key role in supporting and enforcing the overall VRE security, managing components providing authentication, role management, access control, authorization, delegation, auditing, etc. From consideration of the IB security model, and extending that to the VRE environment, it is clear that there are 3 different system “zones” to be handled – the core VRE machines running the portal, the front-end machines providing

access user access, and the back-end machines providing external (to the VRE) resources. Similarly, there are two categories of software that have to be considered, the core VRE components and the 3rd party resources being exploited. While only the core VRE systems and components are directly under the control of the project, security can only be assured if suitable mechanisms and facilities are made available to (and possibly mandated for) the 3rd party components. Many of these components will have their own security model – for example, in the IB project, there is a need to closely interact with the NGS infrastructure, which has an existing extensive security model and infrastructure. Interfacing different security models so that the “join” does not fatally compromise both participating systems is challenging, and more so if a “delegation chain” is required to span the join – such as where authorization from users in one security domain is required so that a component in a second security domain can access controlled resources in the second (or a third) domain.

Though still to be fully designed, the VRE security will be based on the IB security model modified as necessary following an analysis of the VRE-specific assets to be protected and the likely threat vectors. It is anticipated that some sort of token-based mechanisms built around the concept of dynamic virtual organizations will be needed

## 7. Conclusion

This paper outlines the current approaches being undertaken to build a Virtual Research Environment around the tools already being developed by the Integrative Biology project, the current usage scenarios anticipated by this environment and the technical architecture for such an environment. The IB-VRE development is being undertaken within the wide context of a number of VRE developments by Oxford University and the JISC, and it is intended that the IB-VRE will work closely with other VRE developments on common themes and tools. Whilst the IB-VRE is still at an early phase in its development, it offers many exciting development both in supporting Integrative Biology research as well as the more generic research lifecycle.

## 8. References

- [1] JISC VRE Programme, [http://www.jisc.ac.uk/index.cfm?name=programme\\_vre](http://www.jisc.ac.uk/index.cfm?name=programme_vre)

[2] Integrative Biology,  
<http://www.integrativebiology.ox.ac.uk/>

[3] Gavaghan et al, Towards a Grid Infrastructure to Support Integrative Approaches to Biological Research. Philosophical Transactions of the Royal Society of London Series A. Theme Issue on Scientific Grid Computing.

[4] Damian Mac Randal, David Gavaghan, David Boyd, Sharon Lloyd, Andrew Simpson and Lakshmi Sastry. Integrative Biology - Exploiting e-Science to Combat Fatal Diseases. ERCIM News No. 60 January 2005, Special Theme: Biomedical Informatics.  
[http://www.ercim.org/publication/Ercim\\_News/enw60/mac\\_randal.html](http://www.ercim.org/publication/Ercim_News/enw60/mac_randal.html)

[5] Ackermann, E. (1994a). Direct and Mediated Experience: Their Role in Learning. In LEWIS, R. et MENDELSON, P., (eds.), Lessons from Learning. North-Holland, Amsterdam.

[6] Kindberg, T. (1996). WWW: Department of computer science, University of London, also Proceedings of The International Workshop on CSCW and the Web. URL:  
<http://www.dcs.qmw.ac.uk/research/distrib/Mushroom/CSCWWeb.html>

[7] WSRP, [http://www.oasis-open.org/committees/tc\\_home.php?wg\\_abbrev=wsrp](http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsrp)

[8] JSR-168,  
<http://www.jcp.org/en/jsr/detail?id=168>

[9] OGCE, <http://www.collab-ogce.org/nmi/index.jsp>