Integrative Biology VRE

Work Package 2: Initial Analysis Report

Matthew Mascord

Marina Jirotka

Clint Sieunarine

Intended Audience

The purpose of this report is to give an account of the preliminary findings of the research process analysis work undertaken to date within the Integrative Biology VRE (IBVRE) project. The main intended audience is the Integrative Biology Research Consortium who, through the Integrative Biology Project Board are invited to consider the recommendations given in the report and comment on the suggested future course of work for the IBVRE project. Other potentially interested parties include the JISC who, through the VRE Programme Manager are monitoring the progress of this project, the VRE development community at large and the University of Oxford who are intending to use the outcomes of the VRE projects as an input to the development of their inter-disciplinary e-Research strategy.

Issued: 25 Nov 2005

1 Executive Summary

The JISC-funded IBVRE project is developing a large-scale Virtual Research Environment for the Integrative Biology (IB) research consortium. This globally-distributed, inter-disciplinary team is engaged in constructing large-scale software systems to simulate biological behaviour at a variety of spatial and temporal scales, in order to further understanding and hence tackle two of the UK's biggest killers: heart disease and cancer.

This report outlines the preliminary findings of the IBVRE project's work to analyse existing research processes across the IB research consortium in order to establish a set of prioritised high level requirements for the VRE. The methodology used was qualitative in nature and comprised a combination of unstructured, one to one interviews with individual researchers and a focus group at the Integrative Biology project workshop held in Oxford at the end of September 2005.

Future requirements work will concentrate on adopting a phased iterative approach to design elicitation based on the high level requirements identified as highest priority. Within this there is the potential to use naturalistic observation techniques on the heart and cancer modellers.

1.1 Conclusions

1. Highest on the users' priority list for the VRE is support for day-to-day activities rather than activities that occur only once in a funding or research cycle.

2. For the in silico experimentalists, reproducibility of both simulations and figures is critically important but can sometimes be problematic due to limitations in the software tools used and varying experimental practices. Providing a secure, centralised repository of in silico experiments will allow experiments to be trivially reproducible by others, encourage and support best practice, and aid the training of new researchers.

3. The cancer modellers will greatly benefit from learning best practice from the heart modellers – this should be elicited from the heart modellers in future work, documented within the VRE and structures provided within the VRE to explicitly support & encourage it.

4. The VRE should provide a publicly accessible `front-door' demonstration of the core IB infrastructure.

5. Face to face contact is crucially important to initiate, build and maintain good working relationships between collaborators. However for day to day discussions between established collaborators the need shifts towards being able to collaboratively share and manipulate the object (e.g. movie, image, equation, or diagram) under discussion over an audio link – face to face contact becomes much less important.

6. The ability to capture and store collaborative discussions has considerable potential to support the training of new researchers.

1.2 Recommendations

The recommendations of the IBVRE requirements team are as follows (many of these refer to a list of numbered high level requirements given in section 6):

1. The IBVRE team should identify a subset of research groups within the consortium with which to pilot the VRE. Within these groups, active involvement in the VRE design process should be sought at the

    a. technical level, to ensure the VRE system links up to local research process support infrastructure; and at the

    b. scientific level to help steer the development of the user interface.

2. The IBVRE project should pilot the following off-the-shelf technologies with members of the groups identified in recommendation 1:

    a. Anoto [2] digital pen & paper technology for users of paper-based lab books and mathematical modellers (req 4)

    b. Digital whiteboard technology to allow geographically distributed mathematical modelling (req 3)

    c. Collaborative movie annotation software (e.g. Vannotea [1]), as a partial solution to req 2. Users would be able to point to areas of the visualisation movie (figure) but not rotate or otherwise alter the figure.

3. In order of priority, the IBVRE team should carry out the following development activities:

    a. In conjunction with the IB development team, develop a user interface to the secure repository for in silico experiments being developed as part of the IB project (req 1).

    b. Develop a repository for visualisation sessions and cross-links to the integrated information environment (reqs 2 & 5).

    c. Develop a metadata repository (or repositories) that captures the higher level research context within which in silico experiments are carried out (req 5); for the IBVRE team to work closely with the IB development team to ensure the repository links to (or is integrated with), the repositories developed as part of IB work package 4 (data management).

    d. Develop a simulation software issue tracking tool providing traceability to affected in silico experiments (req 6).

    e. Develop a 'linked agenda' meeting scheduling tool (req 9).

    f. Develop a basic facility to upload Anoto paper notes and digital whiteboard sessions to the VRE research context repository (reqs 3 & 4).

4. The IBVRE team should carry out the following integration activities:

    a. Integrate a third-party calendar tool into the VRE (req 10).

    b. Integrate an RSS aggregator tool with keyword filtering to provide a solution to req 7.

    c. Integrate local research process support infrastructure (i.e. individual group wikis & chat facilities) (req 8).

5. The IBVRE team should further establish requirements in the following areas:

    a. The requirements for a more comprehensive solution to provide secure storage of digital whiteboard and Anoto digital paper notes reqs 3 & 4.

6. The IBVRE team should investigate and monitor developments in the following areas:

    a. Use of Personal Access Grids and their integration with VREs (req 11).

    b. Biological data repositories, with a view for their future integration into the VRE (req 12).

7. Due to their perceived low priority, the IBVRE team should not address the following areas unless third-party tools already exist that would be trivial to integrate into the VRE:

    a. Identification of funding area (req 13)

b. Matchmaking (req 14)

c. Proposal writing (req 15)

2 Introduction

The JISC-funded IBVRE project is 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.

2.1 Integrative Biology

Integrative approaches to biology are rapidly evolving and characterised by the attempt to understand biological systems through the construction of large-scale software systems that simulate biological behaviour at a variety of spatial and temporal scales. These *in silico* experiments, as they are known, increasingly demand larger and more powerful resources, both in terms of data storage and computation; the primary aim of the IB project is to meet these needs by constructing a Grid infrastructure to provide tailored, seamless access to these vital facilities.

The partners of the IB project - the IB research consortium - initially represented collaboration between 6 leading UK Universities (Oxford, Nottingham, Leeds, UCL, Birmingham and Sheffield), the University of Auckland, CCLRC and IBM. Interest in the project is such that it has already been expanded to include six additional experimental groups at universities in the US (Tulane, UCSD and UCLA), Canada (Calgary) and Europe (Graz and Utrecht). Researchers within these groups are drawn from a wide range of disciplines, including computer science, mathematics, medical engineering, biophysics, biochemistry, physiology, genetics and several areas of clinical medicine. While the heart modelers are globally distributed, the cancer modelers are all based within UK universities (Oxford, Nottingham and Birmingham) and coordinated centrally as the Tumor Task Force (TTF) by Professor Helen Byrne at the University of Nottingham. The TTF's current aim is to build a multi-scale model of colorectal cancer.

2.2 Main Aims

Whilst work on the IB project has focused on supporting the core scientific workflow of an *in silico* experimentalist (generating, moving, processing, and visualising data on the Grid), the IB Virtual Research Environment (IBVRE) project by contrast, aims to extend this baseline infrastructure to create a usable, comprehensive, and integrated online environment supporting all aspects of the consortium's research processes, from identification of research area all the way through to dissemination and provision of training to the next generation of researchers entering this inter-disciplinary field.

2.3 Current Status

Initial work within the IBVRE project has focused on two main workpackages: (WP3) to create a robust, and flexible infrastructure to allow bespoke and third-party IBVRE tools to be integrated and co-located within a portal framework, and (WP2) to perform an initial analysis of existing Integrative Biology research processes to determine areas where improvement is both desirable and feasible within an integrated environment. This report aims to provide a summary of the WP3 research process analysis work undertaken to date, its preliminary findings, and give a set of recommendations for areas where the VRE development should initially focus.

3 Methodology

To reflect the intended, iterative approach to development, work within WP3 extends almost for the duration of the project. With such a wide remit, the initial activities described in this report aim to scope and prioritise development in order that the first release of the VRE is immediately useful to end users to encourage them to take ownership and steer future development.

The overall approach was qualitative in nature, comprising a set of one to one interviews to determine existing individual research processes and a focus group concentrating more on establishing a set of prioritised high level requirements. Choice of participants was initially guided by the recommendations of the Integrative Biology project manager, Sharon Lloyd but was limited by who was available - the main aim was to ensure the majority of research groups within the consortium were represented. A total of nine researchers and eight research groups had varying levels of involvement in this study:

Participant	**Heart/Cancer Modelling**	**Affiliations**	**1:1 Interview**	**Focus Group**
Dr Blanca Rodriguez	Heart	Computing Laboratory, University of Oxford Computational Cardiac Electrophysiology Laboratory, Tulane University	Yes	No
Dr Jon Whiteley	Heart & Cancer	Computing Laboratory, University of Oxford	Yes	No
Professor Helen Byrne	Cancer	School of Mathematical Sciences, University of Nottingham	Yes	Yes
Dr Chris Bradley	Heart	University Laboratory of Physiology, University of Oxford The Bioengineering Institute, University of Auckland	Yes	No
Dr Carina Edwards	Cancer	Oxford Centre for Industrial and Applied Mathematics	Yes	No
Professor Natalia Trayanova	Heart	Computational Cardiac Electrophysiology Laboratory, Tulane University	No	Yes
Rob Blake	Heart	Computational Cardiac Electrophysiology Laboratory, Tulane University	No	Yes
Dr James Eason	Heart	Virtual Heart Lab, Washington and Lee University Computational Cardiac Electrophysiology Laboratory, Tulane University	No	Yes
Dr Gernot Plank	Heart	Institute of Biophysics, University of Graz	No	Yes

Interviews have also been arranged with Dr Richard Clayton (University of Sheffield) and Professor Philip Maini (University of Oxford) for November 2005, and findings from these interviews fed into the next phase of the analysis work.

3.1 Interviews

The exploratory nature of this study meant that it was neither going to be straightforward nor desirable to construct a very rigid, structured interview format. It was felt that the interviews should be informal, unstructured and open ended to allow the team to probe for further information where necessary. Whilst a flexible technique, it is recognised that this approach does demand more of the interviewer to ensure the interaction stays on track. It was therefore decided to establish a basic interview framework to agree the rules of engagement. The framework had two main parts:

1. To ask the interviewee to provide an account of the day in her life focusing on at least the following aspects:

    a. The higher level research cycle i.e. from identification of research area, and funding stream all the way through to dissemination and provision of training.

    b. The core scientific or experimental workflow i.e. the activities carried out to perform the research or scientific experiments.

    c. The wider context in which the research is conducted e.g. other research projects, administrative or teaching duties.

    d. The artefacts used in this context e.g. workstation, lab book, whiteboard, PDA etc.

2. To probe for information on ways in which collaboration is used to perform the activities described in (1) and the various facets of this collaboration including:

    a. Whether it is synchronous or asynchronous.

    b. Distributed or co-located.

    c. Public or private.

    d. Level of awareness of others required e.g. does the collaboration require face to face contact or is voice contact enough.

    e. Technologies used e.g. phone, whiteboard, email, video conferencing.

    f. Whether there are any issues or barriers to collaboration.

    g. How knowledge is distributed across the different actors in the collaboration.

All interviews were, with permission, taped using a MiniDisc recorder, and where possible carried out by two people: one asking the questions, the other taking notes and monitoring the recording equipment. Post-interview the notes and audio recording were used to produce a short research profile, which was sent to the interviewee for approval.

3.2 Focus Group

Fortunately, within the timescales of this workpackage, the Integrative Biology project had scheduled their yearly project workshop involving nearly all the researchers within the consortium. A focus group devoted to the IBVRE was organised as a breakout session within this and attracted representation from the Graz, Tulane, Washington and Lee, and Nottingham groups. The aim of the focus group was similar in scope to the interviews but the limited time available meant that some sort of structure was introduced to ensure the most efficient use of time. Thus, the focus group was organised into two parts:

1. an open-ended discussion structured around the higher level research life-cycle to elicit high level requirements i.e. a set of candidate tools for integration into the VRE; and

2. a short session devoted to putting the established set of requirements in a priority order.

The focus group was taped, again with permission, a transcript was produced and a summary of the main points distributed to the group for approval.

3.3 Analysis of Findings

This preliminary analysis involved going through the research profiles and focus group transcription identifying common threads and notable idiosyncrasies. The main criteria for selection was whether the research process had relevance in terms of informing the high level design of the VRE; this was largely based on the requirements team's past experiences in specifying, designing and developing online research support systems.

3.4 Future Work

Future work within this workpackage will concentrate on adopting a phased iterative approach to design elicitation based on the high level requirements identified as highest priority. Within this there is the potential to use naturalistic observation techniques on the heart and cancer modellers. The aim is to secure active involvement from users within this at two levels:

- Scientific – in silico experimentalists to steer user interface development
- Technical – those responsible for implementing technical solutions within the research groups to ensure effective linkage to technical solutions already in use locally at research group sites.

The work will also aim to determine best practice guidelines from the heart modellers, looking at them in relation to the cancer modellers.

4 Findings

The findings are structured around the generic research 'life cycle' although there is often significant overlap between the activities within this and they do not always occur in the sequence presented. Those sections marked in *italics* are considered directly relevant in informing the high level design of the VRE.

4.1 Identification of research area

An individual researcher will pursue research avenues developed both on their own initiative as well as brought to them by fellow collaborators. For the latter category, the researcher relies heavily on their contacts in the field. These new research ideas can arise from a variety of sources, for instance from an individual's curiosity, spontaneously during a group discussion, or motivated by a new funding opportunity. Some research ideas are small enough to be realisable within the remit of an existing project whereas others will require a significant amount of effort and additional funding. The point is that the research cycle is not always linked to the funding cycle.

The flexibility to pursue new avenues of research depends to some extent on the discipline. In domains such as Mathematics where research is often carried out by individuals working alone, there is usually the possibility to change direction and impulsively pursue any interesting avenue that presents itself, however in domains requiring large teams, there is likely to be a project plan and any change to this will be subject to some form of authorisation. As mathematicians are used to working on their own, a major challenge for the TTF is getting those involved to work more closely together e.g. by telling each other what they are doing and getting consensus from the group about future directions.

Usually the research idea is kept private until the results of the research are published in a journal or conference. Data is also kept private although this is less important since it is useless without knowledge of what it relates to or how to interpret it. Research groups within the consortium compete as well as collaborate.

From this initial study, it emerges that identification of research area is the activity where it is least clear what the needs are in terms of the VRE, apart from generic tools to facilitate collaboration, and the making of new contacts.

4.2 Identification of funding source

Assuming the research idea is not motivated by a new funding opportunity, the next step is identifying an appropriate funding source. In this area, the focus group revealed a large variation of needs. Since the number of available funding agencies tends to decrease as amounts sought increase, those seeking larger amounts of funding (e.g. to fund an entire lab for a number of years), tended to know where to go to seek funding and so didn't really see the need for tools to help with this. By contrast, those seeking smaller amounts (e.g. to fund a single PhD student), tended to have greater needs in terms of looking for funding sources because of the larger variety of agencies. *The globally distributed nature of the IB consortium led to a further requirement: the ability to search funding agencies globally for joint international funding opportunities.*

Since writing a proposal for funding is a relatively time consuming exercise, those engaged in initiating this usually perform some form of risk assessment exercise before deciding to go ahead. *Factors to consider in this exercise include proposal rejection rate and the make up of the pool of reviewers. Any funding agency search tool would therefore ideally include such information in its search results*. Whether or not this information is available, it appears that an important step in the process is always a phone conversation with the programme manager to get a feel for the likely success of any proposal.

4.3 Identification of collaborators

The need to identify collaborators differed between the heart and cancer modelling communities. As the multi-scale modelling of colorectal cancer is relatively new, there is the pressing need to identify experimentalists who can provide data and help steer the development of models. This involves a lot of searching and cold-calling people to identify enthusiastic people in experimental areas who have an appreciation of the benefits of the maths.

The heart modelling community by contrast have already developed comprehensive models and simulation codes; the focus is, for the most part, on using these existing codes to perform in silico experiments rather than build new models and codes from scratch, and so there is less of a need to identify experimentalists. An exception to this is the Wellcome Trust funded Heart Physiome project which is looking at building a more detailed model of the human heart including in all its processes and intricacies, based on real data from physiologists.

Participants at the focus group felt that any matchmaking tool to help identify people would probably help as a first step but that key to any successful collaboration was the trust between the people involved. This trust is made up of many components and can only realistically be established through face to face meetings. For instance, a researcher needs to make sure that they can work with the other person, be comfortable with their style, be confident that they will deliver on their promises, and at a personal level, that they will 'get on'.

4.4 Proposal Writing

Proposal writing tends to be collaborative, often with a single person, usually the principal investigator leading by producing an outline and delegating sections to individual collaborators. One potential tool put to the participants in the focus group was some form of repository of proposal templates for each funding agency. Whilst no one specifically felt this would be of great benefit, the notion of a tool to help increase the success of the proposal was developed. *The idea was that this tool would, for each funding agency provide a list of keywords and emphases they are looking for. It was felt that any such tool would be useful as a check, and also to help less experienced people.*

In common with the funding search tool, this type of tool would be of most benefit for those who tend to seek smaller amounts from a larger variety of agencies rather than those seeking larger amounts from a handful of agencies.

4.5 Literature Review

Those engaged in building new biological models from scratch need to do a lot of reading of biology journals to inform the design of the models. Unfortunately the wealth of information, especially in fields such as cancer can make this very time consuming. One of the mathematicians interviewed subscribed to alerting services that sent table of contents of relevant journals periodically, but found that it was impossible to scan everything. *A service that could alert a reader to works containing specific keywords could cut down this workload considerably.*

Another aspect of resource discovery is obtaining parameter values for the models. The TTF had used the IB wiki to produce a repository for all relevant data and source papers; interest was expressed in the ability to automate this.

4.6 Project Management

The remit of project management includes a myriad of different activities e.g. developing the project plan, monitoring day to day activities, managing the website, internal/external communication, and organising events. At Tulane, the principal tool used to manage projects for the past 2-3 years is the wiki. The wiki's inherent flexibility allows them to use it to fulfil a variety of functions:

- As a repository for datasets and movies.
- As a library of PDF papers that anyone reads.
- As a repository of all science-related questions and answers that come up.
- As a blog for individual researchers.

As well as the wiki, an internet chatroom is also used extensively at Tulane for more technical related topics. Its success can be measured by the fact that the majority of researchers at the lab monitor it peripherally throughout the day. As the chats are archived and forwarded to a shared, searchable mail account, anyone asking a question that has already come up can be referred to the chat archive. In this way, the chatroom becomes an effective knowledge & experience capture tool.

As the cancer modellers are just getting started, and do not yet have a large multi-scale model, their main needs are in tying the project together. The research team building the multi-scale model comprises a number of PhD students working on individual models and postdoctoral research assistants who are responsible for linking the models together – a real management challenge is ensuring that each PhD student has an independent project not overlapping with other projects (to protect their PhD), yet there is enough coverage of the different spatial and temporal scales within the multi-scalar model of colorectal cancer. *A key finding from the focus group was that the cancer modellers would benefit greatly from learning best practice from the heart modellers. Eliciting this from the heart modellers, documenting it and designing the VRE in such a way as to support it is the highest priority requirement for the cancer modellers.*

Another finding from the focus group was that any kind of shared calendar or scheduling system is unlikely to be of use since it requires buy in from everyone in order for it to succeed. Such as system could however be used to notify the group about relevant up coming event.

4.7 Scientific Workflow

The notion of scientific workflow for the purposes of this report is defined as the day to day activities that an individual researcher is engaged in and directly relating to their research agenda. It is this area that exhibits the greatest variation amongst those interviewed; whilst it is possible to think of the Integrative Biology research process in the whole, each individual researcher will be engaged in one specific part of this with its own unique set of needs. *As yet it is unclear whether the responsibility for supporting scientific workflow lies in the remit of the IB or the IBVRE project.*

4.7.1 Mathematical modeller

The mathematically modellers interviewed described the mathematical modelling of biological systems as a combination of face to face brainstorming sessions with biologists and analytical work carried out individually. The aim of the brainstorming sessions are to:

- help the mathematician gain an understanding of the biology by asking lots of questions; and
- produce models that are meaningful yet tractable mathematically.

A typical session will involve the mathematician and the biologist both in front of a whiteboard writing diagrams and equations, depending on the stage into the collaboration. The analytical work involves:

- writing small-scale simulation codes in MatLab or Fortran to understand how the model behaves; and
- pen and paper work to obtain a greater insight into the biology through a search for certain asymptotic limits for which it is possible to write down explicit solutions, these can show more directly how changes to parameters affect the observed behaviour.

Model building is driven largely by intuition. The issue with intuition is that often lots of people have different intuitions based on mathematical models they've seen; experimentalists also have very strong intuitions over what they think the key biological processes are: anyone's intuition can be wrong. The mathematics can sometimes test the biologist's intuition. The multi-scale model of colorectal cancer being constructed by the TTF will be a plug and play model, with many alternative sub-models designed to answer different types of questions. There will however, be a default e.g. to be able to see a tumour developing in the colon.

Whilst most simulations in MatLab or Fortran are currently small-scale and performed as proof of concept exercises for mathematical models, there will come a point when the TTF will start to scale the simulations up to a level where large-scale Grid facilities will be necessary. The feeling came through that access to Grid facilities before this would be useful to get a feel for them, to ease the transition so the modellers will be able to hit the ground running. The main barrier to this appeared to be the impression that it requires training in order to use it, the lack of documentation, and just not being able to see where to start. There was the feeling of not wanting to put someone out (in terms of training) in order to play with it, if there is no real need for it yet. *The requirement here in terms of the VRE is that, if it is to encourage use of the Grid, should provide a simple 'front-door', to the base-level IB infrastructure, perhaps not initially requiring a digital certificate. If these services are required, the procedure for obtaining a Grid certificate should be well documented within the VRE, and simple to follow.*

4.7.2 Numerical Algorithm Developer

An intermediate role between the mathematical modeller and the developer of the simulation software is the developer of the algorithms that will be used in the eventual simulation software. Dr Jon Whiteley, a lecturer at the Oxford University Computing Laboratory identifies (and sometimes develops) the fastest and most accurate numerical algorithms for use in simulation software. Dr Whiteley described his scientific workflow as comprising two phases:

- a literature review to identify new numerical algorithms; and
- development of simple C++ programs to assess their accuracy and speed.

Dr Whiteley is not attempting to produce *the* fastest codes in this process, he is merely trying to identify if the algorithm has the potential to run in the fastest time (if proper software engineering and optimisation techniques were utilised). Accuracy in this context means that the algorithm accurately simulates the mathematical model.

4.7.3 Simulation Software Developer

A huge variety of development platforms and methodologies are used by those engaged in developing simulation software systems. The main programming languages used by those interviewed were Fortran, C++ and MatLab. Developers can be divided into three categories: those enhancing and maintaining an existing large-scale simulation software system, those starting to develop new large-scale simulation software systems, and those prototyping new, small-scale simulation codes as a proof of concept exercise for new mathematical models or algorithms.

As might be expected, the developers of the larger systems had a more pressing need for comprehensive software development support tools. Like any technical development endeavour, source code control and issue tracking are high up on the priority list, but not just to ensure the consistent delivery of high quality software: *for the in silico experimental process, disciplined version control and issue tracking are critical to ensure both the reproducibility of experiments and the traceability required to track bugs to software versions and software versions to affected experiments*. These needs have been tackled in a variety of ways by the groups in the consortium but a more tailored solution specific to the in silico experimental process could have a very positive impact on the integrative biology research process both within the IB project and the community at large.

A further need that came out through an informal conversation with one of the developers was help writing the simulation code itself. It turns out that sometimes the most time consuming and error prone parts of development is writing the code to parse input files: *if the generation of the declarations and parsing code could be generated automatically, this effort would be cut down substantially, freeing the developer to work on the vital numerical code.*

4.7.4 In silico Experimentalist

In silico experiments are carried out in two phases: simulation and visualisation. If computational steering is used, then these two activities overlap to some extent because there is the opportunity to stop the simulation mid-run, visualise any partial results, and re-start the simulation with new input parameters.

At Tulane, the main simulation executable is launched from scripts written by the in silico experimentalist. In the main, the bulk of these scripts are devoted to data management: transforming input files into a form that the simulation executable can understand. Sometimes the script will launch the simulation executable iteratively e.g. to perform multiple pacing shocks on the virtual heart. The development and maintenance of the simulation executable (memfem) is handled centrally by Rob Blake.

Every new in silico experiment at Tulane potentially has a need for changes to input files, scripts and the simulation executable itself, either to model a new type of behaviour or capture new data. It is very difficult to predict what these changes will be in advance and unrealistic to imagine a scenario where a simulation system could be developed that could accommodate any new requirement as a configuration change. *Therefore, to ensure reproducibility of experiments, version management of input files, simulation software and scripts is critically important.*

At Tulane and Washington and Lee, generating movies and images to visualise the results of experiments is performed within the CoolGraphics software [4]. Given a set of output files from a simulation, reproducing a figure can be problematic because CoolGraphics cannot be invoked programmatically from a script. The only way to reproduce a figure is to manually interact with the CoolGraphics desktop application following instructions written by the original experimentalist. The alternative graphics package, Meshalyzer [3] does have this capability and so Tulane are keen to move to it, the factor currently preventing this is its lack of support for certain graphical capabilities.

To help resolve problems with the reproducibility of simulations and figures, the focus group decided that there was a pressing need for a repository of in silico experiments. This repository would capture everything that is needed both to replicate both the simulation and any visualisation (movies, images) of the experiment, in an end to end, bullet-proof manner.

Most of the in silico experimentalists interviewed kept some form of a lab book whether paper-based or a digital equivalent. One of the computational biologists at the Oxford University Computing Laboratory had used a paper-based lab book since her PhD to record thoughts, input parameters to experiments, reading lists, figure cut-outs from papers, and results. Associated with the use of the lab book were a number of rules such as never removing a page in case there is a need to go back to it. This researcher said that she would always prefer pen and paper over an online equivalent but that it would be a disaster to lose her book; she said that she sometimes photocopies pages from her lab book if people need them.

Such accounts point to the potential of Anoto [2] digital pen & paper technology, almost identical in look and feel to their traditional counterparts except that everything written is digitally captured within the pen for subsequent upload to secure storage. An experimentalist preferring pen & paper would be able to continue to use a medium they are comfortable with yet be safe in the knowledge that anything that is written down will be safely and securely backed up online, as well as being more accessible to themselves and their collaborators.

4.8 Real-time Communication

The needs for real-time i.e. synchronous communication varied again by discipline and by the stage into the collaboration. The heart modellers used it mainly to discuss results over the phone; whereas the cancer modellers would have frequent face to face meetings to flesh out models with experimentalists and mathematicians.

4.8.1 Cancer Modelling

In mathematics, there is nearly always the need for an equation or picture that can be shared and annotated by participants in the discussion: for one to one discussions this can be a piece of paper, for larger discussions a whiteboard or blackboard is the main tool used. *One of the mathematicians interviewed felt that for one to one discussions, a useful addition to face to face meetings might be some form of digital whiteboard solution, larger meetings being harder to coordinate 'virtually'.* Such technology, if it were sufficiently robust and easy to use would make meetings easier to organise and spontaneous, would open up the number of people that can participate, and would save time since a whole day can sometimes be consumed travelling for a relatively short meeting. All this could help the mathematicians feel closer even if they are not, geographically. The only proviso to this was that any whiteboard system would have to be natural to use i.e. pen-based and should not be expected to replace face to face meetings entirely since face to face contact is the only realistic way to develop and maintain the trust necessary for a successful working relationship.

4.8.2 Heart Modelling

When remotely discussing a movie relating to a heart modelling experiment there are two options: either both parties can download the same movie, or both parties can download the applicable dataset and configure the visualisation software to reproduce the movie directly. In the latter case, there is then the flexibility to rotate or otherwise alter the figure through the visualisation software; the disadvantages are that the dataset is nearly always larger in size than the movie and that there are sometimes issues configuring visualisation software to reproduce visualisation movies exactly.

One interviewee reported that often when discussing a result, there is the need to swap between different movies, and it is difficult to know in advance which movie or data set will be required. Where both parties do not have access to the same sets of data, the discussion is disrupted whilst the relevant data is downloaded and visualisation software configured**.** *In this case, some form of dataset and movie repository with a caching tool that ensures each party participating in the collaboration always has local access to exactly the same datasets and movies could help alleviate this.*

Another common problem was the inability to point to the part of the movie that is being discussed; as one participant put it at the focus group, 'sometimes it's hard to describe exactly which wavefront you're supposed to be looking at.'. *This introduces the need for some kind of pointing device that can be shared between participants. As a further requirement, there was strong interest in the ability to record the discussion synchronised with the playing of the movie and any pointer movements.* This would be useful

both for reference and also to give to a postgraduate student, particularly since the student is rarely involved in the discussion that originally started the collaboration.

This could be achieved either by using one of the existing movie annotation tools or by adding this functionality to the visualisation software itself. One example of a tool that has been developed specifically to address this type of problem is the Vannotea [1] prototype developed by Dr Jane Hunter's group as part of the FilmEd project at the University of Queensland, Australia. The feature most relevant to IBVRE is its `Collaborative video analysis and discussion' facilities, these provide capabilities such as synchronous playback, collaborative drawing, and multiple mouse pointers.

4.9 Dissemination

In the IB project it is important to disseminate results in the life sciences community to engage experimentalists who can help steer the development of the models; for a physical scientist this can be challenging because the life science journals require a different slant, and use different languages. Both conferences and journals were used by those interviewed. Conferences can either serve as publicity vehicles (in which case a larger conference would be targeted), or as a forum in which to do collaborative work (usually the smaller conferences or workshops). Conference papers are often the presentation of the work in progress, with perhaps several conference papers for a single project, whereas the journal paper will be longer and the final word on the subject.

Paper writing practices varied amongst those interviewed but tended to follow the same pattern as proposal writing; usually a single person, the first author, would lead by producing an outline which is agreed either by all the co-authors or just the first author and the supervisor. Depending on who is leading, the first draft will either be written by the first author, or co-authors will complete individual sections. The first author will be responsible for editing the final draft for submission to the journal or conference.

A variety of tools are used to prepare papers, mathematicians tending to use LaTeX; in silico experimentalists preferring Word or OpenOffice. CVS is sometimes used in conjunction with LaTeX to manage versions of papers. If Word is used, then Microsoft Tracking is often used to track changes to papers. Some participants used BibTeX (a bibliographic database system compatible with LaTeX) to manage and present reference lists in papers.

Those at the focus group felt that paper writing is an inherently asynchronous activity, the ability to write papers synchronously with others, would not really be of benefit because writing papers was something best done alone to enable concentration and the ability to iteratively write and re-write.

Another important aspect to dissemination is what happens to all the workings out, that lead to the final paper. This can be important material since it can contain many previously disregarded avenues that may be ripe for future exploitation. The mathematicians interviewed said that in the main these workings out tended to be written on pieces of papers, grouped together in folders and stored in a filing cabinet. One of the mathematicians expressed some concern about this since there is no backup. When at the paper writing stage, a lot of the maths gets written up in LaTeX documents but a lot of the material is not in a form that can be easily typed in: lots of scribbles, annotations, arrows, etc. *There was strong interest in the idea of the Anoto [2] digital pen & paper technology as a great way to back up notes, and make it easier to access material when away at conferences and events.*

4.10 Training

The Integrative Biology project has provided funded for 10 PhD students to enter the Oxford University Life Sciences Interface Doctoral Training Centre (DTC). This is a partially taught postgraduate course designed to train graduates in the physical sciences in biology and advanced mathematical and computational skills, equipping them with the tools required to enter research in this inter-disciplinary field. DTC students are provided with access to the Oxford University Bodington Virtual Learning Environment (VLE) system and this is used primarily to hold all the lecture notes and worksheets that they ever see. During the course DTC students complete mini-projects, an example of this would be reproducing a result from a paper by coding and other mathematical modelling techniques. In the second part of the course, the students undertake their substantive DPhil research project at the life sciences interface within one of the

application areas, based within the research groups of their principal supervisor at their home institution. Informal discussions with the director of the Doctoral Training Centre, Professor David Gavaghan revealed that many of the DTC students had developed strong bonds whilst at Oxford and that many of them were missing each other. *To help address this, a VRE might provide tailored access to video conferencing facilities such as Personal Access Grids to help maintain these relationships.*

At Tulane and Oxford, wikis are used extensively to manage the training process. At the Oxford University Computing Laboratory, one researcher described how she uses wikis to train PhD students in the use of Linux and the relevant tools. The student is pointed at a set of documentation and 'How to' notes in the wiki. The wiki is preferred to email because items only have to be uploaded once and the wiki is a public record of everything that has been given to them - with email there is the risk that material is lost. At Tulane, each PhD student uses the wiki as a blog and they use this to post interim results. This saves time in progress meetings because the supervisor can pre-screen results and come to the students with questions about these results rather than spending time asking what the student has been doing; it also helps the supervisor to see whether the student has understood what they have been asked to do in the first place. Not all of the researchers interviewed shared this belief in the benefits of wikis and blogs as training tools. One researcher interviewed felt that students are more likely to read an email than a wiki entry, that there is a risk that the student might look at the wrong part of the wiki, and that an email or a knock on the door is more efficient.

It was pointed out at the focus group, that the requirement for an experiment repository (outlined in section 4.7.4) to capture everything that is needed to reproduce both the simulation and any figures (movies & images) resulting from these simulations could also aid the training process of new students. Students would be able to reproduce any in silico experiment performed by any experimentalist in a simple point-and-click manner, perhaps with the opportunity to change a few of the parameters to see what happens.

5 Conclusions

1. Highest on the users' priority list for the VRE is support for day-to-day activities rather than activities that occur only once in a funding or research cycle.

2. For the in silico experimentalists, reproducibility of both simulations and figures is critically important but can sometimes be problematic due to limitations in the software tools used and varying experimental practices. Providing a secure, centralised repository of in silico experiments will allow experiments to be trivially reproducible by others, encourage and support best practice, and aid the training of new researchers.

3. The cancer modellers will greatly benefit from learning best practice from the heart modellers – this should be elicited from the heart modellers in future work, documented within the VRE and structures provided within the VRE to explicitly support & encourage it.

4. The VRE should provide a publicly accessible `front-door' demonstration of the core IB infrastructure.

5. Face to face contact is crucially important to initiate, build and maintain good working relationships between collaborators. However for day to day discussions between established collaborators the need shifts towards being able to collaboratively share and manipulate the object (e.g. movie, image, equation, or diagram) under discussion over an audio link – face to face contact becomes much less important.

6. The ability to capture and store collaborative discussions has considerable potential to support the training of new researchers.

6 Summary of High Level Requirements

The following table presents the set of high level requirements elicited from the interviews and the focus group. Each requirement is assigned a high, medium, or low priority categorisation according to the priorities established in the focus group and, (for those requirements not explicitly prioritised at the focus group) a judgement based on perceived interest, frequency of occurrence and how closely they address the needs of day to day research activities. Those activities explicitly prioritised at the focus group are marked with an asterix (*).

As the remit of the VRE project is, roughly speaking, a superset of the IB project's remit (addressing the needs of the IB research process as a whole rather than focusing solely on performing in silico experiments on the Grid), requirements identified through this initial analysis exercise are inevitably going to overlap with some of the main aims of the IB project. To clarify cases of overlap, a further analysis was performed to establish whether for each requirement the responsibility for developing a solution lies with the IB or IBVRE project. In all cases, the assumption is that the IBVRE project will be responsible for providing the user interface for any solutions developed.

Req #	Name	Description	Priority	IB or VRE?
1	In Silico Experiment Repository*	The ability to access and trivially reproduce – subject to access constraints defined by the experimentalist - any in silico experiment and figures relating to the experiment performed by researchers within the IB consortium. Everything relating to the experiment would be captured e.g. script versions, simulation software version, input files, movie/image creation scripts. Optionally, the ability to change minimal parameters associated with such experiments.	High	IB[1]
2	Collaborative Visualisation*	The ability for two or more geographically distributed in silico experimentalists to synchronously replay visualisation movies and point to areas of the movie while discussing it over an audio link; the ability to capture and securely store online such a visualisation session (audio synchronised with the movie replay and pointer movements).	High	IB + VRE[2]

[1] This VRE requirement is addressed by a combination of the IB project's work packages 1, 2, 3, 4 and 5 (executable management, workflow management, job submission, data management, and visualisation); the VRE would be responsible for providing the user interface, and through VRE requirement 5, cross-links to related entities e.g. papers containing figures produced by the experiment, to provide the *context* within which the experiment is performed.

[2] Visualisation software development is under the remit of the IB WP5 (visualisation), storage of visualisation sessions is not under the original remit of the IB WP4 (data management) because it was not in the original scope of IB, the responsibility for this functionality therefore lies within the scope of the VRE; providing a portal interface to such a repository is also in the remit of the VRE project. VRE is responsible for implementing collaborative viewing tools based on the visualization and movie creation components developed in IB as well as trialing any third-party tools that address this requirement.

3	Distributed Mathematical Modelling	The ability for two or more geographically distributed mathematical modellers to share a 'virtual whiteboard' allowing participants to collaboratively draw and annotate mathematical equations and pictures whilst discussing the mathematics over an audio link; the ability to capture and securely store online such a session - both the audio & whiteboard markings.	High	VRE
4	Backup and Access to Paper Based Notes	To allow an mathematician or experimentalist to continue to use paper and pen for individual mathematical workings or as a log book, but for the contents of this work to be automatically captured digitally and securely stored online; providing a backup and easier access when away from the office.	High	VRE
5	Integrated Information Environment	The ability to create and store online arbitrary associations between artefacts relating to the in silico experimental process e.g. experiments linking to figures in papers, visualisation sessions, wiki pages, chat archives.	High	VRE
6	Simulation Issue Tracking	The ability to capture and manage issues/bugs associated with simulation software versions – for experimentalists to be alerted to these issues where they may affect any previous experiments carried out.	High	VRE
7	Paper Notification	The facility for modellers to be notified of articles in selected journals matching specific keywords.	High	VRE
8	Local Research Support Infrastructure Integration*	The ability to access existing research group support and management infrastructure (e.g. wiki & chat repositories), through the VRE.	Medium	VRE
9	Rapid Access to Visualisation Movies	The ability for the participants of a visualisation session to identify required resources (e.g. visualisation movies & datasets) and pre-download them to local machines in advance of the session.	Medium	VRE
10	Online Newsletter	Tools to help writing the TTF newsletter e.g. possible online version of this with upcoming events calendar, progress reports, publications & presentations.	Medium	VRE
11	Real time video communication	The ability to communicate in real time with others over a video link.	Medium	VRE
12	Biological Data	The ability to construct repositories of relevant biological data through the VRE, with linkages to existing biological data repositories.	Low	VRE
13	Funding Identification*	The ability to search for international funding opportunities in heart and cancer modelling through the VRE; to be able to retrieve information on proposal rejection rate for the applicable funding agency, and the make up of the pool of reviewers.	Low	VRE
14	Collaborator Matchmaking*	The ability to search for potential collaborators, and make initial contact with them, through the VRE.	Low	VRE

15	Proposal Writing*	The ability to capture and store information likely to increase the success of proposals being funded e.g. by holding sets of keywords or approaches that individual funding agencies are looking for, as well as approaches that the funding agencies will reject.	Low	VRE

7 Recommendations

The recommendations of the IBVRE requirements team are as follows (many of these refer to a list of numbered high level requirements given in section 6):

1. The IBVRE team should identify a subset of research groups within the consortium with which to pilot the VRE. Within these groups, active involvement in the VRE design process should be sought at the

    a. technical level, to ensure the VRE system links up to local research process support infrastructure; and at the

    b. scientific level to help steer the development of the user interface.

2. The IBVRE project should pilot the following off-the-shelf technologies with members of the groups identified in recommendation 1:

    a. Anoto [2] digital pen & paper technology for users of paper-based lab books and mathematical modellers (req 4)

    b. Digital whiteboard technology to allow geographically distributed mathematical modelling (req 3)

    c. Collaborative movie annotation software (e.g. Vannotea [1]), as a partial solution to req 2. Users would be able to point to areas of the visualisation movie (figure) but not rotate or otherwise alter the figure.

3. In order of priority, the IBVRE team should carry out the following development activities:

    a. In conjunction with the IB development team, develop a user interface to the secure repository for in silico experiments being developed as part of the IB project (req 1).

    b. Develop a repository for visualisation sessions and cross-links to the integrated information environment (reqs 2 & 5).

    c. Develop a metadata repository (or repositories) that captures the higher level research context within which in silico experiments are carried out (req 5); for the IBVRE team to work closely with the IB development team to ensure the repository links to (or is integrated with), the repositories developed as part of IB work package 4 (data management).

    d. Develop a simulation software issue tracking tool providing traceability to affected in silico experiments (req 6).

    e. Develop a 'linked agenda' meeting scheduling tool (req 9).

    f. Develop a basic facility to upload Anoto paper notes and digital whiteboard sessions to the VRE research context repository (reqs 3 & 4).

4. The IBVRE team should carry out the following integration activities:

    a. Integrate a third-party calendar tool into the VRE (req 10).

b. Integrate an RSS aggregator tool with keyword filtering to provide a solution to req 7.

c. Integrate local research process support infrastructure (i.e. individual group wikis & chat facilities) (req 8).

5. The IBVRE team should further establish requirements in the following areas:

   a. The requirements for a more comprehensive solution to provide secure storage of digital whiteboard and Anoto digital paper notes reqs 3 & 4.

6. The IBVRE team should investigate and monitor developments in the following areas:

   a. Use of Personal Access Grids and their integration with VREs (req 11).

   b. Biological data repositories, with a view for their future integration into the VRE (req 12).

7. Due to their perceived low priority, the IBVRE team should not address the following areas unless third-party tools already exist that would be trivial to integrate into the VRE:

   a. Identification of funding area (req 13)

   b. Matchmaking (req 14)

   c. Proposal writing (req 15)

8 References

[1] Vannotea <http://metadata.net/filmed/prototypes.html>

[2] Anoto <http://www.anoto.com/>

[3] Meshalyzer, Dr Edward Vigmond, <http://www.enel.ucalgary.ca/~vigmond/>

[4] CoolGraphics, Dr James Eason, <http://theinferno.wlu.edu/VirtualHeart/index.php/Main/LabSoftware>

9 Acknowledgements

The VRE requirements team wishes to thank Blanca Rodriguez, Jonathan Whiteley, Helen Byrne, Chris Bradley, Carina Edwards, James Eason, Rob Blake, Natalia Trayanova, and Gernot Plank for sparing their time to be interviewed or for participating in the focus group. Thanks also go to David Gavaghan, Matthew Dovey, Andrew Simpson, Sharon Lloyd, Damian Mac Randal, and Geoff Williams (members of the IBVRE project management board) for their input and approval of the recommendations contained within this report.