JISC Information Environment Portal Activity: supporting the Needs of e-Research. – Scenarios, Use Cases and Reference Models –

Rob Allan
CCLRC e-Science Centre, Daresbury Laboratory,
Daresbury, Warrington WA4 4AD

Rob Crouchley
Centre for e-Science, C Floor Bowland Annexe, Lancaster University, Lancaster LA1 4YT

Caroline Ingram
CSI Consultancy Ltd., 42 Coquet Terrace, Newcastle upon Tyne NE6 5LE

Contact e-Mail: r.j.allan@dl.ac.uk, r.crouchley@lancs.ac.uk,
caroline@csiconsultancy.co.uk

May 18, 2009

Abstract

Scenarios, Use Cases and Reference Models are analysed for a number of e-Research projects. This analysis will inform our recommendations of how the JISC Information Environment and its portal interfaces could be enhanced to support the needs of e-Research, in particular for resource discovery.

This report should be read together with others which form part of the same study, particularly the interim and final reports and our vision for a global Information Environment.

© **CCLRC 2006.** Neither the CCLRC e-Science Centre nor its collaborators accept any responsibility for loss or damage arising from the use of information contained in any of their reports or in any communication about their tests or investigations.

Contents

1 Introduction

Scenarios, Use Cases and Reference Models are analysed for a number of e-Research projects. This analysis will inform our recommendations of how the JISC Information Environment and its portal interfaces could be enhanced to support the needs of e-Research, in particular for resource discovery.

2 Analysis of e-Research Projects

We follow the methodology of the STORE Project [5]. Some of the scenarios and use cases are taken from the DigiRep Wiki [6] in cases where they seem relevant to the IE and e-Research.

We attempt to adhere to the terminology and methodology of the emerging e-Framework for Education and Research, see `http://www.e-framework.org`.

2.1 What is a SCENARIO?

A scenario is a brief narrative, or story, that describes the hypothetical use of a system or process. In one or more paragraphs, a scenario:

- Tells who is using the system/ process and what they are trying to accomplish;
- Provides a realistic, fictional account of a user's constraints: when and where they are working, why they are using the system/ process, and what they need it to do for them;
- Describes any relevant aspects of the context in which the user is working with the system/ process, including what information the user has on hand when beginning to use it;
- Gives the user a fictional name, but it also identifies the user's role, such as student, faculty member, staff, or general public;
- Indicates what the user regards as a successful outcome of using the system/ process.

2.2 What is a USE CASE?

In software engineering, a use case is a technique for capturing the potential requirements of a new system or software change. Each use case provides one or more scenarios that convey how the system should interact with the end user or another system to achieve a specific business goal. Use cases typically avoid technical jargon, preferring instead the language of the end user or domain (subject) expert.

Use cases which have informed the development of the Information Environment are documented [20]. They include the basic cases for: enter; survey and discover; detail; use record; request, authorise, access; use resource.

2.3 What is a REFERENCE MODEL?

The following is from a presentation by Bill Olivier (JISC Development Director). It was given early in the process of defining the language of the e-Framework, so may need updating.

In the context of a catalogue of Service Components [**?**] (the "Wall of Bricks"), a Reference Model:

- Selects the Component Services
- Orchestrates and/ or Choreographs them (Orchestration: several services working together for a user) (Choreography: several users/ organisations working together) [9]

We therefore need to identify:

- The Human-level Tasks and Workflows(which may be improved or redesigned in the process)
- Show how these relate to the Service infrastructure
- Identify the Service-level Workflows/ Processes

We shall not have time to do this thouroughly for more than one example but attempt to illustrate the process and draw conclusions for the rest of the study.

2.4 Worked Example from Social Science Research

A scenario is short story that describes the functions in a context. To illustrate this we do not need need use cases as this is far too detailed for the broad picture we want to present. In fact most of the DigiRep examples were scenarios, not use cases. We show below how a scenario might inform the necessary components of the IE architecture and services. Similar analysis can be done for the series of scenarios identified in Section A. This set should be widened based on other reviews and studies which have been completed recently. Our worked example is based on e-Social Science and its differences with other disciplines are dedscribed in Section 3.

A possible Social Science Researcher's (SSR) scenario is.

1. Suppose we have a researcher (SSR) who could access all of the Archived Data Sets and those used in every social research publication in their research field and decide on the most appropriate data for their needs, without having to spend days reading through coding schedules and questionnaires;
2. Suppose SSR could automatically re-estimate all the models others have used on these data sets, and see what happens if you drop or add new variables to the analysis;
3. Suppose SSR could quickly formulate (check the identification etc.) and estimate any new models or combinations of existing models you thought might be relevant;

4. Suppose SSR could re-do this across multiple datasets;
5. Suppose SSR could match your research questions to information held in existing digital resources. Search for new explanations;
6. Suppose SSR could integrate multiple sources of data and text to help to fill in missing data and ideas.

Services (or steps) required in this scenario include:

- search publications and archived data sets
- select and download appropriate data matching a particular research need
- re-construct previously used models
- re-compute models on these data sets
- re-compute models on these data sets with different parameter choices
- compare results
- creat new models or combine existing ones
- repeat analysis across multiple datasets
- match research questions to digitally-stored information
- integrate multiple data and text sources to identify missing data and ideas

What does this imply for the architecture? Well the original IE architecture diagram from Andy Powell [10], Figure 1 is missing a few key functions/ elements.

To illustrate the missing elements we can look at the mapping between the various Suppositions of this Scenario and the elements of the AP's diagram.

1. At first inspection the notion of content as used by AP and archived data sets may be different, so we need to be clear that MIMAS, EDINA and the various Archives, ESDS Data archive as well as the data archives of online journals that contain copies of the data sets used by journal authors, etc. So perhaps instead of content we should use the phrase DR in our use of these diagrams and make it explicit. Also its not clear what is the presentation layer that SSR would use. It could be a project VRE, or a Social Science gateway, that is cross connected to the DRs. It could be just a browser, but that would lack functionality, so we need to add these to the front row. This part of the scenario highlights the need to link data sets to publications [1]. See Figure 2.
2. A this stage, the problem with Figure 2 is that it does not contain any computational facilities, so these have been added. Not sure what symbol to use for the actual kit, or whether this is not needed as its implied [2]. See Figure 3.

[1] One JISC-fuded project investigating this is CLADDIER [12]. JISC is also funding links to data, e.g. on MIMAS and EDINA via the JCSR in the GEMS project.

[2] JISC is funding computational facilities for the support of research via JCSR, e.g. the National Grid Service [13]

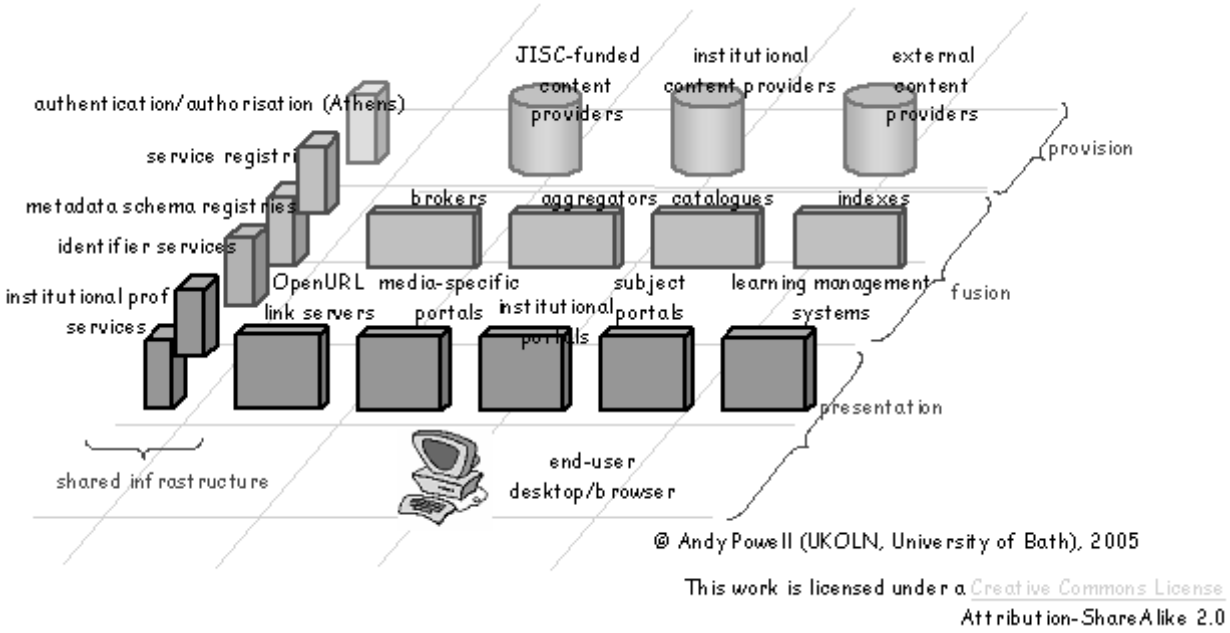

Figure 1: IE Architecture version 1

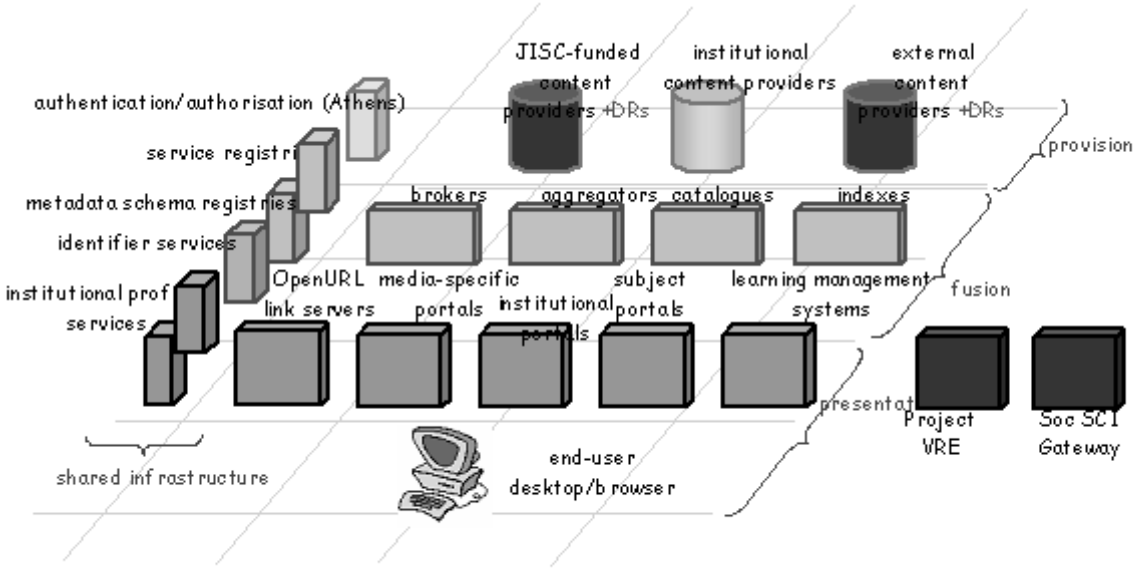

Figure 2: IE Architecture version 2

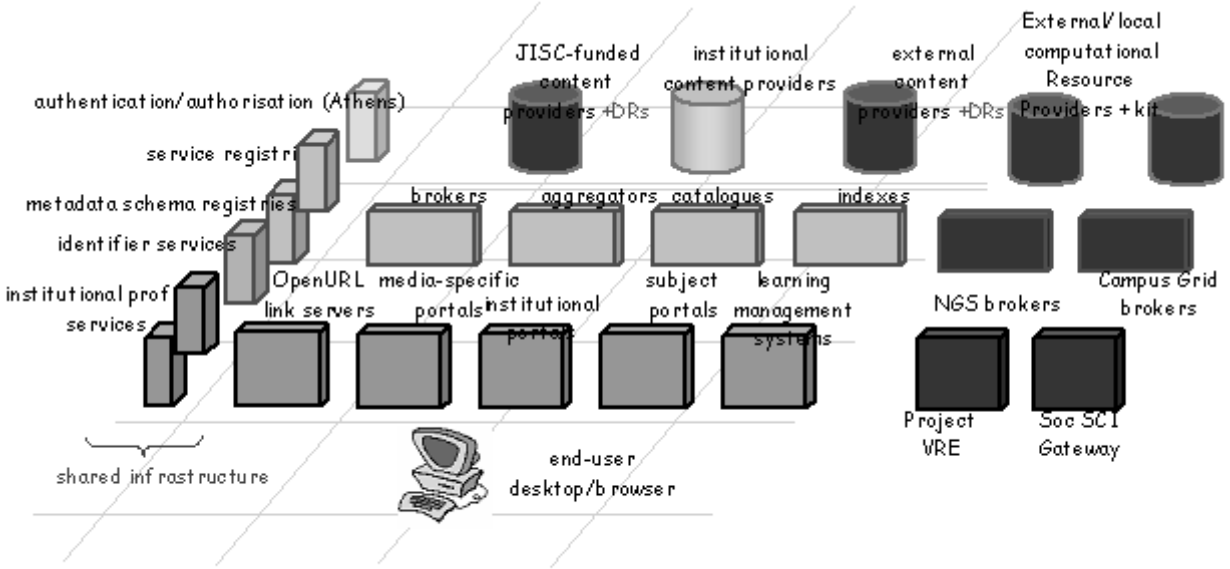

Figure 3: IE Architecture version 3

3. This stage requires the creation of new tools, which run on the computational facilities, so we need to be clear that its not enough to provide the physical infrastructure interconnections there will be a bunch of new software tools and middleware. This will have implications for the choice of service components.

4. This stage is going to jointly use the DRs and the computational facilities.

5. This stage is going to use the journal literature (External Content), but will require the use of new content harvesting and synthesising services/ tools.

6. This stage requires new tools as well that extend the functionality of those in stage 5 but add the data to the mix.

It therefore seems that it is not sufficient for us to simply add elements to the original IE Architecture diagram we also need to look at the "brick wall" diagrams coming from the ELF and e-Framework activities `http://www.e-framework.org` to see what elements we need for our scenarios.

We analyse the Social Science Research (SSR) scenario again with this in mind:

1. This uses the highlighted elements;

2. This uses the highlighted elements (4 new ones);

The other stages can use the same services.

[some figures?]

But its not clear yet what toolkits we need to fulfil this scenario. In the ELF they had, see below, so we need something like this building.

Area	Scope	Relevant Framework Toolkits
Assessment	The creation, execution and recording of electronic assessments.	APIS (Assessment Provision through Interoperable Segments)
Resource Discovery	Search and discovery of e-learning content across a range of repositories.	D+ (Brokerage for Deep and Distributed e-Learning Resources Discovery) MDC (Middleware for Distributed Cognition) and JAFER
Learning design and Sequencing	The design, construction and execution and exchange of learning activities.	SLeD (Service-based Learning Design) and Coppercore ISIS (Integrating Simple Sequencing)
Enterprise	The sharing and management of student information.	SWEET.net toolkit CETIS IMS Enterprise Toolkit *
Personal Development Planning	The creation, recording and sharing of PDP and ePortfolio information.	WS4RL toolkit (PDP)
Portal Services	Services for the display and embedding of external data in institutional web-portals	PSE (Portal Services Embedder)

3 Disciplinary Differences in e-Research

A lecture was given by Christine Borgman, 14 June 2005. This is available on the Web: `http://webcast.oii.ox.ac.uk/?view=Webcast&ID=20050614_73`. The following notes are based on this lecture and relevant to the analysis of e-Research scenarios. See also [2].

e-Research is a collective term for the various initiatives on e-Science, e-Social Science, e-Humanities, and cyberinfrastructure. e-Research refers to distributed, collaborative, information-intensive forms of inquiry. The overall aim is to do faster, better, and different interdisciplinary research (and scholarship) across the university, as summed up by Prof. Tony Hey, then head of the U.K. e-Science programme.

e-Social Science research currently is organized into two themes: (1) research and development of technology, tools, and data sources to support collaborative social science research; and (2) social study of e-Research. e-Research in all disciplines will depend upon the generation, analysis, visualization, management, and curation of data and documents, and upon access to those resources.

Interdisciplinary research will depend upon sharing data within and between communities. Decades of research in information studies and in socio-technical systems has shown that disciplines vary greatly in their use of data and documents, in their local or distributed access to information resources, and in their degree of collaboration. Understanding more about the use of information is essential to the construction of an information infrastructure to facilitate research.

The talk surveyed behavioral, social, political, economic, technical, and institutional information issues that vary between disciplines and suggested research that is needed to inform e-Social Science.

There could be said to be a current data deluge in many areas of research – academics are producing

more data than they can easily use. This implies a need for enhanced organisation and information maagement, which can be greatly facilitated through the use of meta-data.

Data and documents don't have to be in the same repository, brokers using the meta-data can tie repositories together.

We talk about building an "information infrastructure", a term that has technical, social and political connotations: people, technology, tools and services are required to provide a distributed collaborative environment.

The cyberinfrastructure layered model is roughly as follows:

Information and knowledge layer: content/ digital libraries, scientific DRs.
Middleware services layer
ITC infrastructure processors, memory, network layer.

There is also an applications space in the cyberinfrastructure for building user interfaces and tools

It is relatively easy to provide resource discovery, but harder to build an infrastructure for information – looking at information in a particular context (i.e. between a social scientist and a technologist there would be different views of information and data).

An infrastructure for information would contain:

- Documents, most previous work on publications – scholarly, theses, government documents and pre-prints/ reprints
- Data – experiments, observations, surveys, interviews, geospatial, includes census data;
- Composite objects – growing importance, difficult to offer – studies of habitat environmental data, observations, field notes, analytical reports on political behaviour, virtual reality models of archaeological sites, visualisations.

Bringing these together will be hard. There is an NSF report on composite objects, e.g. opinion surveys, census etc, how can it be brought together as a package. `find reference`

Some examples of why these are needed are seen in the other scenarios in Section A, in particular the Silchester Roman Dig VRE. Other examples are: (2) NASA has posted the data and various images that the public can use, interpreted from the data through visualisation tools `find reference`. (3) UCLA centre for network sensing – multiple data used (biology e.g.) macro level data in weather to macro level in soil information `find reference`. (4) 3D building VR models of archaeological sites, including timing data, historical data to construct things that are no longer there from certain periods `find reference`.

Research questions

Who's value chains are best served by the information infrastructure? Everything of use to some people – but how do you decide what data to include with which documents? Could consider: What's likely to be reused? What's likely to be share between multiple users/ students? How might sharing be different between and in fields? How is context provided?

3.1 Literature

Christine Borgman has done research on information seeking and using literature between different disciplines.

Commonalities:

• Scholarly publications are input to research, there is an expectation that what's published will input to future research
• Peer review gives an ability to assess quality, replicate, verify the research
• Access to bibliographic records transcends disciplines
• Active curation to make things available long term – push for repositories for scholarly communication – libraries lease access to electronic resources, do not have long term control
• Linking adds context whether tying docs together (citation refs) or linking data to docs.
• Some data to show items in Repositories get more links `???`

Disciplinary differences

• Collaboration: co-authorship is likely to result from collaborative research – the bigger the team the more likely there is external funding
• More books = more sole authors
• Time frame for usefulness of documents: in some sciences if things aren't immediately cited then they are no longer useful. In cell biology can be as little as 2 months, in systematic biology 2 centuries, in humanities can be 20 years.

Subjects with short document half lives are commonly the ones building repositories and getting involved in technologies for sharing documents since it is more important to them to get rapid access; long half life projects are associated with more book publishing

Quick wins are possible, though not easy, for document intensive research; i.e. improving access to journal articles benefits everyone in subjects with short half lives who are more likely to be getting immediate benefits as vast proportion of material online and has immediate access. Other subjects want documents going back several decades (monographic literature) will not be gaining as much – they need automation of catalogues and indexes so that they can locate what's available offline.

3.2 Data

There is a similar analysis for scientific data.

Commonalities

With regards to data there are disciplinary commonalities

• Individuals are generating a lot of data (especially historians)
• Individuals are also using others' data
• Sharing including multi institutional projects
• Metadata within institutional fields
• Data visualisation

• IPR agreements complex in all fields
• Quality control standards and practices a problem – no peer review for data and repositories. Data is used in proof ???

Differences

Even more disciplinary differences with data than with documents

• There are a great number and large variety of data types – hard to inventory what's out there for certain subjects (art history in comparison with chemistry – in the latter for example robots can be used to identify all instances of an element, in the former letters, x-rays of painting, and documentary histories may need to be accessed.
• Ability to identify other sources: natural sciences have best odds for finding other data sources of relevance, social sciences, medium and humanities a low likelihood.
• Agreement on representation varies greatly – in some subjects, there is no agreement for what something is called (art history again, spellings may be different, or language of use different). Even in natural sciences this can be problematic, for instance there are projects at Cambridge and Daresbury investigating this for computational chemistry
• Metadata formats

- General discovery metadata
- Field and instrument specific micro level metadata

• Sensitivity, sharing, especially medical records and social sciences interviews, but also for environmental, e.g. endangered animals and plant locations
• Data mining, lack of consistency; i.e. in chemistry, easy to find, and relatively consistent, but also very valuable for pharmaceutical industry if made freely available. Social surveys hard to resell; stock market data, most valuable for the first 20 mins in which the data is available. Fish tracking, can sell for first 24 hours, but weather researchers will find it be useful for weeks
• Obligations to deposit the data vary. In most subjects individuals keep their own. Wellcome Trust requires deposition within 1 year; others in US – varies greatly, so as disciplines work together ???
• There are differences between security in institutions; different practices
• Agreement between data repositories, people may want to deposit but have nowhere to put their own data, e.g. recordings of language research. Journals in which interpretations are placed may not want to store recordings too.

Issues for sharing data:

- Context and interpretation – scholarly publications, for centuries end result – once in print could destroy data in past
- Take data and place in data repository, removing the context, often numeric data requiring metadata,
- Although some data is a little more self describing, tacit knowledge is stripped away – this can result in data loss so will not gain from sharing further.

- Essential to develop trust and reciprocity – want to replicate and compare results for usefulness in sharing

Lots of incentives not to share data

- Promotion based on publication; no benefit to publishing data
- Concern about data deposit and others publishing based on data – hence need embargos – can be metric on embargo, collection time vs. allowing access
- Loss of tacit knowledge, loss of control of IP, is one of the biggest issues – current research reconstructing a public domain around data by creating a legal contract

In the UK these issues are exaccerbated by the RAE (Research Assessment Exercise) which raises the issues to a departmental/ institutional level.

3.3 Summary

The goals of e-Research are to build a value chain of information- intensive multi-disciplinary research. Building an infrastructure for information is complicated as it will need to deal with context issues for use and reuse

Documents: are common across disciplines, practice in production and curation and linking; but differences in degree of collaboration and time frame of usefulness

Data: some things in common, generating large volumes, IPR, quality control; but differences greater between disciplines for data issues, types, sources, representation, metadata formats, sensitivity, value, whether DRs exist and can be funded

Context and interpretation: publications provide context for many repositories. This is one of arguments for linking data and docs, as docs can provide data that's been stripped away when data put into DR if tied together

Many incentives to share are for sharing for the common good, altruistic, and funding agencies are appealing to public good as public money is spent – but incentives not to share are real parts of scholarship – controlling access to data, being the first to publish, cultural reasons, ownership and priority issues.

So in considering social shaping of information infrastructure we need to know more about context, use and re-use and sharing, access to information in the larger infrastructure will depend on behaviour, trust and policy issues rather than focussing on technology issues.

4 Other related Work

4.1 NeSC Workshop

A workshop on *User Requirements and Web based Access for e-Research* was held at NeSC on 19/5/06. A technical report has been published [21] and material from the workshop can be found at the URL cited. This workshop comprised of dual presentations from a user and a developer on the following e-Science projects: Discovery.Net; e-Minerals; GEMEDA; myGrid; e-HTPX; RealityGrid; and one US TeraGrid project, GEON. There were also 3 breakout sessions focussing on: user-developer relations; co-modification of solutions; and functionality.

In general, the workshop noted the follow:

- There is a lack of agreement of common practices;
- There is a need for closer interaction between developers and tool builders;
- There is a need for better understanding of successful application/ tool building teams;
- There is a lack of agreement on basic functionality;
- There is a need for common base implementation – and a common base system supported in a production manner;
- Given a common base implementation, there will then be a need for subject specific extensions to the common base system.

Michelle Osman of Discovery.Net noted that *data sharing and integration and support for dynamic, iterative workflow development is a critical functional requirements for their system*, i.e. linking to tools which offer this through a portal where results and data can be searched for and aggregated.

There was a comment/ issue raised in one of the breakout groups on the *Difficulty of aligning competing requirements across user communities with very diverse aims and criteria for success.* This is more evidence (albeit slightly anecdotal) to back up the discussion on disciplinary differences.

Unfortunately there was no more discussion of access to information repositories and only minor mention of publication of data and results from the myGrid and GEON projects. This may indicate a need to make researchers more aware of the possibility and importance of these activities. This is confirmed in the Digital Repository Roadmap [22] *The majority of academics do not know what repositories are nor are they familiar with the issues around new means of dissemination.*

4.2 StORe

StORe: Source to Output Repositories looked at linking repositories of source data with repositories of publications (output) in 7 subjects and carried out interviews with scientists in these subjects to clarify their requirements.

Subject Area	Source Repository	Output Repository
archeology	archeology data services ADS	White Rose IR
astronomy	SuperCOSMOS science archive	Edinburgh Research Archive; arXiv
biochemistry	protein structures database PDB	UCL Eprints
biosciences	UniProt; Genbank	PubMed Central
chemistry	National Crystallography Service	Imperial Eprints
high-energyphysics	Brookhave, CERN	Birmingham Eprints
social policy and political science	UK Data Archive UKDA	LSE Research Articles Online

We note that this table only contains one or two repositories in each entry, there are actually many more.

The principal objective of the survey of seven scientific disciplines was to identify aspects of desired functionality in source and output repositories that would be included in a mechanism for enabling source to output links. The survey found that the proposed two-way link between source (data) repositories and output (publications) repositories was considered useful, but that there were a number of cultural and organisational barriers to the deposit of research data in source repositories. These included concerns over workload, frustration with bureaucratic processes and uncertainty with respect to the protection of intellectual property. Amongst the survey constituents there is also a perception that repositories are inconsistent in terms of their coverage, metadata and formats. An almost universal preference for simple methods of searching was declared, with positive reference made to the Google experience, although more advanced searches are undertaken when specific data sets have been identified. In the organisation of data and when using repositories, self-reliance is more common than recourse to institutional, library or other support, although the need for expert support in the development and maintenance of metadata and other standards is acknowledged. Given the large scale and complexity of many research data generated, together with the idiosyncrasies of the individual software and data platforms used to produce them, it was thought that links from processed data would in certain instances be more useful than links to/from raw data. Nonetheless, a significant degree of consensus was found over the requirements for core metadata; there was also a general agreement with the principle of open access.

In July 2006, the StORe survey work identified the following strategy:

- The pilot middleware could provide a core generic solution capable of accepting discipline-specific add-ons;
- A core and standard platform for metadata can be established to reflect common practices and needs;
- Cross-discipline data requirements must be met for output and source data;
- Different attitudes to data sharing have to be supported by effective validation if repositories are to be accepted and effective;
- Online rather than personal support should meet most user expectations.

They also identified the fololowing additional questions which need to be addressed in building real software:

- What is the real demand for access to source data and what do the researchers want to do with it?
- What improvements to searching does the research community really require?
- How far will the cross-discipline use of source data influence options to standardise across data sets?
- Which source repositories are used or not used, what works and what doesn't?
- Are there common practices between disciplines, are standard metadata systems and schemes in use, and is a generic solution feasible?
- Options for support - is the demand for human mediation or online solutions?

5 Reference Models based on Scenarios and Use Cases

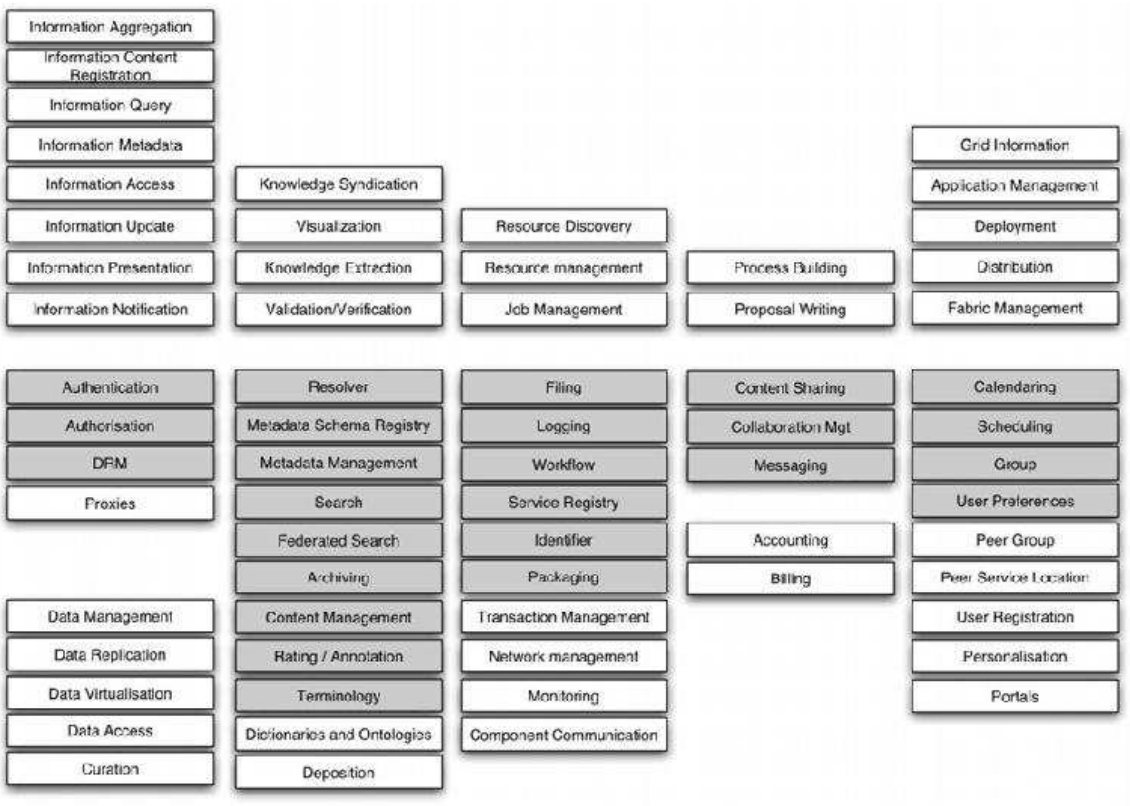

Figure 8 E-Science Framework (JISC JCSR)[6].

Figure 4: Reference Model for e-Research

In this section we present a reference model, similar to that created previously for e-Research and documented by Scott Wilson [1]. This combines the services identified in the various use cases which will be described below in Section A.

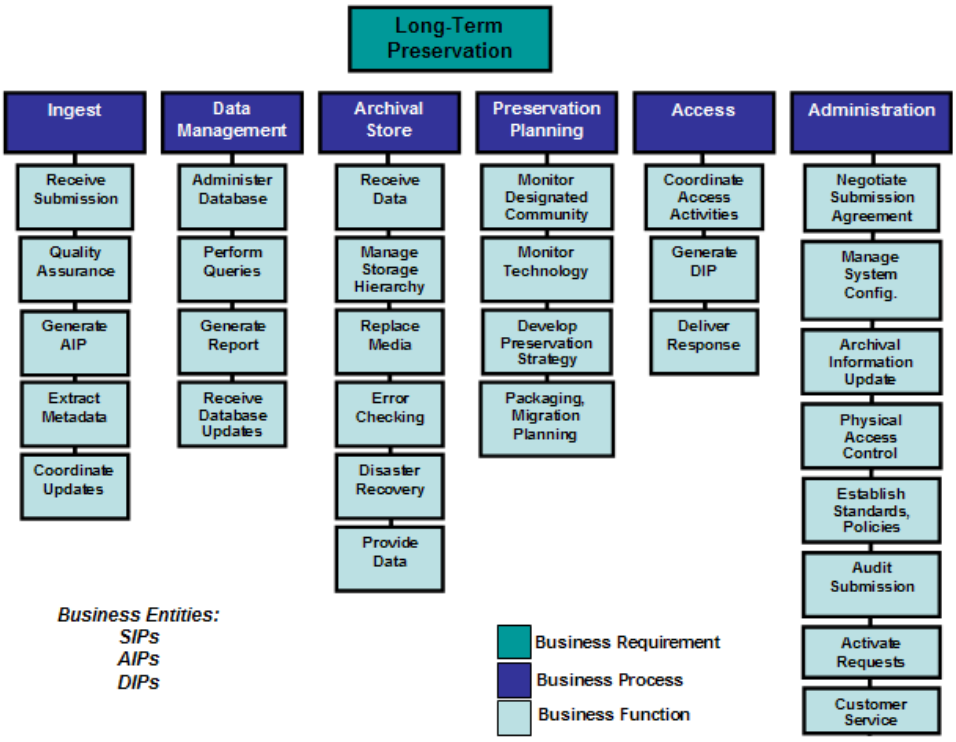

Figure 5: OAIS Reference Model

Another reference model is described by the Open Archival Information System (OAIS) for digital preservation [19]. Figure 5 shows that it is organised as a set of processes representing the overall library business. The Service Framework Group in the USA has similar goals to the e-Framework for Education and Research to which JISC contributes.

A generic research "life cycle" was defined in the Integrative Biology VRE project to include:

1. Identification of research area
2. Identification of funding source
3. Identification of collaborators
4. Proposal writing
5. Literature review
6. Project management
7. Scientific workflow
8. Real time communication
9. Dissemination
10. Training

Another version of a project life cycle was defined in the EVIE project [4].

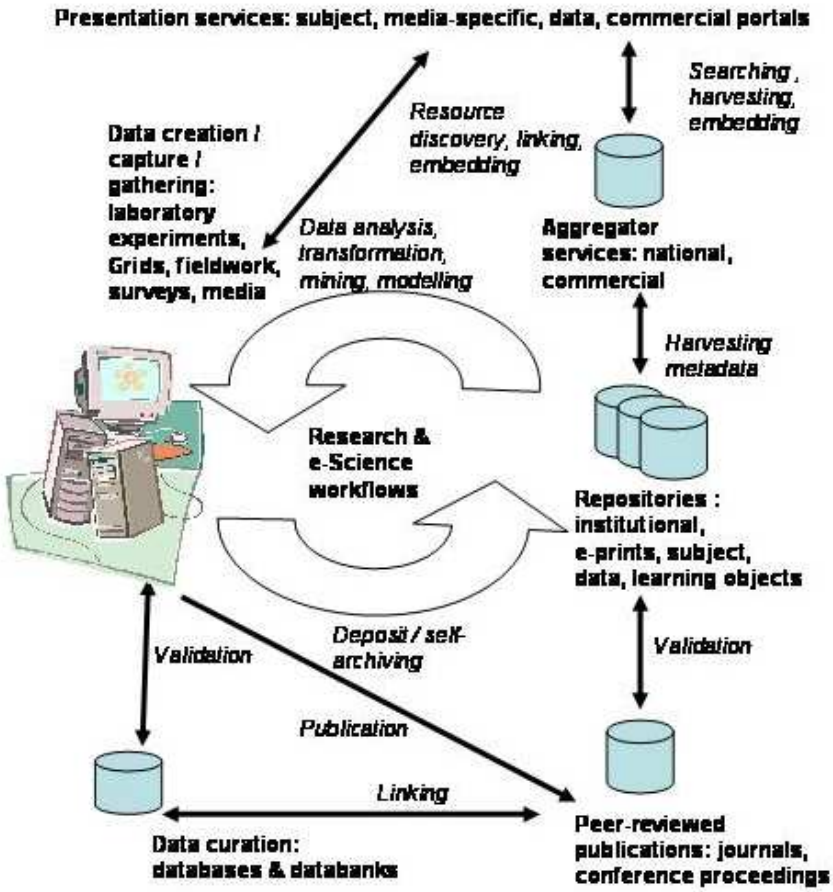

Figure 6: Research Knowledge Cycle addressed in eBank UK

This has been set in pictorial form a number of times, e.g. by the eBank project [3] showing the scholarly knowledge cycle 6, and by Matthew Dovey in discussions of the JISC e-Framework for Education and Research.

Services listed in the use cases have been "genericised" and include:

- distributed query resulting in links to source data
- various query interfaces depending on context (e.g. text, video, etc.)
- consolidate various data types resulting from search
- do context-based query, e.g. on geographical information or based, text or identifier (individual or sample)
- do query based on annotated video material
- extend queries on private information to external sources

- enter and publish data, adding new samples, individuals and relationships
- need standalone and networked tools
- register to use a digital repository
- sign in
- access import/ deposit tool
- compress files to agreed format
- upload to submission area
- repository system does validation/ assertion checks
- confirm copyright status
- notification of errors to be fixed by user
- create and upload metadata
- provide quality measure
- validate metadata and save
- preview and re-edit if necessary
- submit for publication
- review by repository administration team
- team moves submission to public area and sends notification to user
- researcher reads instructions – help service
- select a subject-specific centre
- decide on format for submission
- create submission with necessary forms (meta-data)
- submit to collection using interface
- collection officer assigns id, completes process log, does validation etc.
- upload to executive
- notify submitter and iterate until complete with no errors
- backup and preserve. Make public copy available
- add new publication
- view my publications
- view unsaved entries

- view draft entries
- view entries submitted but not yet approved
- carry out admin functions
- sign out
- devise experimental strategy subject to COSHH regulations
- submit to system including metadata for artefacts and people involved
- set permissions based on identifiers and roles
- recall plan on laboratory control device
- select one or more experimental techniques and appropriate workflows
- lead laboratory worker through experimental workflow with appropriate decision making
- automatic recording of data and metadata, including instrument and software versions
- archive this information (data capture and deposition) for later retrieval
- convert data formats from proprietary to common
- submission of sample labelled with originator and identifier
- delegation of experimental tasks to trained facility staff
- scheduling of experiment
- potentially do remote collaborative working with facility staff
- collection of time-stamped metadata
- assertion service
- recording of digital images
- assessment of quality and sustainability
- data post processing – solution, refinement, report preparation
- archival and curation – e.g. off-site tape store
- search based on artefact identifier to find related information and include in report
- selection of template for report based on research technique (method) used
- choice of journal and report generation in standard journal format
- deposition of report in institutional repository
- create links from report to underlying datasets
- artefact/ sample production and monitoring

- remote visual checking of artefacts using a robot and camera
- image analysis service and feature recognition
- data collection
- labelling using a bar code and categorisation
- safety check at the facility
- shipping service is used to "FedEX" the artefacts
- tracking of artefacts
- maintenance of artefacts, e.g. if they are fragile or require special storage treatment
- scheduling an experiment or other research process
- data collection
- data quality control and automation of dat collection process
- data analysis using various techniques – comparision of results
- deposition – upload of data and meta-data for international researchers to use.
- identification of research area
- identification of funding source
- identification of collaborators
- proposal writing
- literature review
- project management
- real time communication
- dissemination
- help and training
- access from Web browser or Matlab
- incorporating computable models of biological systems into executable simulation codes and installing them on a range of systems;
- job preparation – specifying the input parameters of a simulation problem to be solved;
- selecting the resources to be used and submitting the simulation to run
- monitoring and controlling the simulation during execution;
- computational steering of jobs
- securely managing and curating both input and output datasets;

- creating appropriate metadata information for future reference;
- managed access to data
- simulation subsystem which provides a range of solvers for user-supplied model codes, possibly running in coupled mode
- storage of users files along with associated metadata
- data retrieval
- curation
- visualisal examination of simulation and experimental results
- collaborative visualisation working with remote colleagues
- search publications and archived data sets
- select and download appropriate data matching a particular research need
- re-construct previously used models
- re-compute models on these data sets
- re-compute models on these data sets with different parameter choices
- compare results
- creat new models or combine existing ones
- repeat analysis across multiple datasets
- match research questions to digitally-stored information
- integrate multiple data and text sources to identify missing data and ideas

5.1 Research Reference Models

These have been converted to a number of smaller "Research Reference Models" which are available in an accompanying spreadsheet. They represent various aspects of the research life cycle and are as follows:

RRM1: Computation or other Task – reflect scenario from GROWL, e-Minerals projects

RRM2: Participation in a collborating Peer Group – e-Minerals, Archeology

RRM3: e-Publication, similar to the OAIS "ingest" process – see AHDS, e-Pubs, R4L

RRM4: Content Discovery and delivery, the latter part similar to the OAIS "access" process – ReDRESS, SPP, e-Pubs, OAIster

RRM5: Diamond Light Source

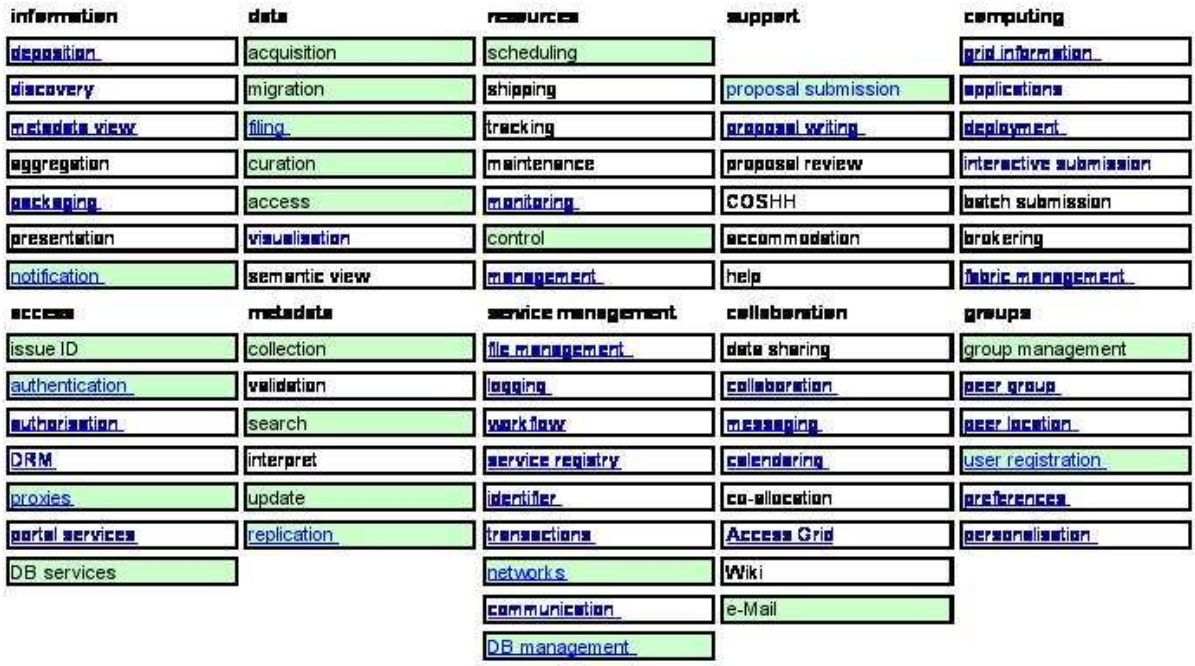

Figure 7: Research Reference Model for the Diamond Light Source

RRM6:

RRM7:

These will be put on-line as progress is made, `http://www.grids.ac.uk/Papers/Classes/framework.html`.

Whilst it is clear that a number of groups worldwide are thinking in similar terms, the exact details of the functionality and its experssion as re-usable services in the implementation of such reference models is still under debate. We will however find in our survey of the functionality and status of the Information Environment [27] that steps are already being taken to translate this into reality. In that paper we shall attempt to identify existing components and gaps where new services and components will be required to meet the needs of e-Research.

5.2 Physical Artefacts

We note that a feature of some e-Research use cases is that physical artefacts have to be handled. This is illustrated by the handling of chemical samples in the R4L and e-HTPX examples. Health and safety (COSHH) regulations and other constraints (e.g. durability of the sample) are constraints. Monitoring, scheduling, tracking, shipping services are required. There are additional constraints on services which manipulate devices such as robots and telescopes.

Services relevant to physical artefacts are outwith the scope of the current study and will not be further considered.

5.3 Workflow

The reference model above lists separate generic services which are of use to the e-Research community. Clearly to benefit from these services they must be linked in workflows which reflect the business processes as outlined in the various use cases. Whilst we present the services as being separate, loosely coupled and idempotent, this may not always be the case and there are likely to be certain dependencies. Services responsible for authentication and authorisation are obvious candidates. Further discussion is however outwith the scope of the current study.

5.4 Portals and User Interfaces

Several of the scenarios below have identified the need for a variety of user interfaces, including portals. Matlab was mentioned as a desktop interface in the IB project and Stata has been mentioned in SSR projects involving statistical analysis. They also identify the need to support both no-line and off-line working. In some cases they identify the need for contextual interfaces, e.g. a search based on annotated video material.

User requirements identify the following functionality that a portal, or indeed any user interface, should have:

- Searching and results lists should integrate bibliographic records with finding aids;
- Keyword searching is preferable, with an ability to narrow and refine the results;
- Compound "advanced" searching is often useful, with limits by date, repository, location, keyword;
- Searching by region/ repository is desirable;
- A summary of all contributing institutions is useful;
- A resource needs a distinctive name;
- It is desirable to be able to bookmark pages, results and search criteria;
- Ability to share search critera and results or post to a list;
- Helpful information on how to contact individual archives and obtain local guidelines for using the material;
- Help and finding aids opening in another window;
- "landing pages" that can be accessed by Google and other general search engines;

This preliminary analysis of user requirements for the interface could be extended and lead to a design study with appropriate stakeholders and focus groups.

6 Conclusions

A simple all-embracing generic use case for "discovery to delivery" in research might be as follows:

```
A researcher wants to carry out a subject-specific search via one or
more portal interfaces and to be able to find relevant publications
and data associated with their studies and to be able to find other
papers which cite them. He/ she may also want to find associated grant
references and appropriate funding opportunities for related work.

The researcher then wants to access and download some of the datasets
and carry out a similar piece of work using a new model, new insight
or adding new data to the previous study. In an experimental study
they might be repeating a recommended procedure on one or more new
samples or applying an improved procedure to a benchmark sample.

The researcher will afterwards discuss and share results with a peer
group, using appropriate personal and group information management
software and will eventually create reports and publish the results
together with related data and model information.
```

We have not completed a full analysis of any of the scenarios and use cases which we have collected. However we have illustrated a methodology which seems to be emerging from various JISC activities. In brief:

- Researchers want access to data and information (e.g. scholarly publications) for a variety of reasons. They want to access all sources in a seamless way and to have a uniform style of presentation;
- They want to use the results of such discovery for a variety of purposes, fusing data and information from multiple sources;
- They want to use previously stored data and also create new data and information from computational or experimental procedures;
- They want to publish new data and information, potentially from personal repositories into public repositories;
- Research Reference Models can be developed based on research processes outlined in the scenarios and use cases;
- these RRMs represent parts of the generic Research Lifecycle;
- RRMs can be realised as Designs using generic service components (this hypothesis is yet to be fully tested);
- The IE Architecture can be extended with additional components to accomodate an implementation of these designs in real Artefacts;
- A range of context-based user interfaces are required to access components in the extended IE architecture;
- Use of the components and services can be facilitated by workflows supporting the research process;
- Many activities worldwide are beginning to implement parts of this overall architecture and we need to integrate with them;
- However, toolkits to support the implentation of most are not there available.

7 Acknowledgments

JISC for funding.

We thank people we spoke to: Colin Nave, Dave Meredith, Mike Gleaves, Catherine Jones, Simon Hodson, Paul Beckett, David Gavaghan, Matthew Mascord, Sharon Lloyd, Graham Klyne, Stephen Lyon, David Zeitlyn, Michael Fischer, Janet Bagg, Simon Coles, Derek Sergeant

References

[1] S. Wilson, K. Blinco and D. Rehak *Service Oriented Frameworks* DEST (Australia), JISC-CETIS (UK) and Industry Canada `http://www.jisc.ac.uk/uploaded_documents/AltilabServiceOrientedFrameworks.pdf`

[2] Christine Borgman *Building a Usable Infrastructure for e-Science: An Information Perspective* Keynote talk at the UK e-Science All Hands Meeting 2005, (Nottingham, UK, 19-23 Sep 2005) `http://www.nesc.ac.uk/talks/ahm2005/keynote1.ppt`

[3] L. Lyon and S.J. Coles *eBank UK: linking research data, learning and scholarly communications* JISC Joint Programmes Meeting, Cambridge, UK, 7-8 July 2005 `http://www.ukoln.ac.uk/projects/ebank-uk/dissemination`

[4] D.M. Sergeant, S. Andrews and A. Farquhar *Embedding a VRE in an Institutional Environment (EVIE). Workpackage 2: User Requirements Analysis* User Requirements Analysis Report (University of Leeds, 2006)

[5] *StORe: the Source to Output Repositories* project funded by JISC and CURL from 1/9/05-31/8/07 has carried out a detailed survey of researcher method and practice in the use and management of digital repositories, and involes the research communities of 7 scientific disciplines. Reports are published in the Edinburgh Research Archive `http://www.era.lib.ed.ac.uk` and also available via the Consortium of University Research Libraries, CURL `http://www.curl.ac.uk`. See `http://www.jiscstore.com/WikiHome`

[6] *DigiRep Wiki* the on-line Wiki resource for the JISC Digital Repositories Programme. It is maintained by UKOLN `http://www.ukoln.ac.uk/repositories/digirep/index/JISC_Digital_Repository_Wiki`

[7] Matthew Mascord, Marina Jirotka and Clint Sieunarine *Integrative Biology VRE, Work Package 2: Initial Analysis Report* `http://www.vre.ox.ac.uk/ibvre/IBVREInitialAnalysisReport.pdf` University of Oxford (November 2005)

[8] Graham Klyne *SakaiVRE User Requirements* `http://wiki.oss-watch.ac.uk/SakaiVre/UserRequirements`

[9] Eric Newcomer and Greg Lomow *Understanding SOA with Web Services* Addison Wesley, 2005

[10] A. Powell *Information Environment Architecture* `http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/`

[11] K. Cole

[12] B. Matthews and C. Jones *Citation, Location And Deposition in Discipline and Institutional Repositories* `http://claddier.badc.ac.uk/trac`

[13] A. Richards

[14] K. Steinbrenner *The Information Architecture Imperative* Educause Research Bulletin vol 2003, issue 2 (ECAR 21/1/03) `http://www.educause.edu/ir/library/pdf/ERB0302.pdf`

[15] D. Spicer and B. Metz *A new Model for Supporting Research at Purdue University* Educause Research Case Study (ECAR) `http://www.educause.edu/LibraryDetailPage/666?ID=ECS0507`

[16] D. Spicer and H. Blustain *Digital Humanities at the Crossroads: University of Virginia* Educause Research Case Study (ECAR) `http://www.educause.edu/LibraryDetailPage/666?ID=ECS0506`

[17] IT Support of Research Subcommittee *e-Research needs Assessment* (University of Iowa, 16/2/06) `URL`

[18] Prof. Mark Greengrass *RePAH: Research in Portals in the Arts and Humanities.* (University of Sheffield 2005) `http://www.ahrcict.rdg.ac.uk/activities/strategy_projects/repah.ppt`

[19] B. Lavoie, G. Henry and L. Dempsey *A Service Framework for Libraries* D-Lib Magazine (July-August 2006) `http://www.dlib.org/dlib/july06/lavoie/07lavoie.html`

[20] A. Powell and L. Lyon *JISC Information Environment Architecture* `http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/functional-model/`

[21] J.M. Schopf and I. Coleman (eds.) *User Requirements and Web base Access for e-Research Workshop* NeSC technical report XXX available from `http://www.nesc.ac.uk/technical_papers/XXX` Workshop presentations available at `http://www.nesc.ac.uk/esi/events/685/`

[22] R. Heery and A. Powell *Digital Repositories Roadmap: looking forward* (Eduserv and UKOLN, 2006) `rep-roadmap-v15.doc`

[23] M.J. Smith it Use Case Compendium of Derived Geospatial Data (GRADE Project, December 2005) `http://www.edina.ac.uk/projects/grade/usecasecompendium.pdf`

[24] R.J. Allan, R. Crouchley and C. Ingram *Scenarios, Use Cases and Reference Models* (CCLRC, June 2006)

[25] R.J. Allan, R. Crouchley and C. Ingram *Comparison of Surveys* (CSI Consultancy, June 2006)

[26] R.J. Allan, R. Crouchley and C. Ingram *Web-based Library and Information Services* (CCLRC, June 2006)

[27] R.J. Allan, R. Crouchley and C. Ingram *The Information Environment and e-Research Portals* (CCLRC, June 2006)

[28] R.J. Allan, R. Crouchley and C. Ingram *Interim Report* (CCLRC, June 2006)

[29] R.J. Allan, R. Crouchley and C. Ingram *A Vision for Portal access to Global Information* (CCLRC, June 2006)

[30] R.J. Allan, R. Crouchley and C. Ingram *e-Research, Portals and Digital Repositories Workshop* [30] Notes from the workshop held at University of Lancaster 6-7/9/06 (CSI Consultancy, September 2006)

[31] R.J. Allan, R. Crouchley and C. Ingram *Final Report* (CCLRC, June 2006)

[32] C.S. Ingram *IE Inventory with Images* (CSI Consultancy, 2005)

A Scenarios and Use Cases

In the following sections we have taken input from a number of projects. We have extracted relevant information in the form of a scenario with accompanying use case and other information. The aim is to reduce these as far as possible to a set of common components which might fit the e-Framework paradigm.

A.1 Anthropology

Name:	Anthropologists and e-Research
Scenario:	Querying datasets composed of disparate data types via user configurable front ends. Data types - text, visual (still and video), audio, numeric, spatial, genealogical, simulation and models, more???
Use Case:	I want to be able to query data contained in dispersed locations and generate usable results with embedded hypertext links to the source material, and I want to be able to do this from a variety of interfaces depending on the kinds of queries I have. For example, I want to be able to generate a genealogical map of a community or lineage and by selecting individuals or sub-sets of the genealogy [1] generate mini reports of data linked to those individuals in field notes, residential data (presumably spatial like GIS), publications, visual (video clips or still images), audio clips and so on. At other times, I may want to query the data via a clickable map which can help me to ask questions about residence and movement of individuals in the area. At yet other times, I may want to watch a video and be able to click on individuals or objects in the video and generate the same sorts of reports. Basically I would like to be able to generate common resource files for all data types and when querying one of the data types to be able to use the same query mechanims to simultaneously query all the other data with compatible resource files. Ultimately, I'd like these to be coded in ways which make it possible for the queries to be expanded to external sources as well (eHRAF, JSTOR, WoK, designated Web sites, but I want to be able to be more selective than "all the Web" searches). Finally, I would like to be able to use the same tools to enter data as well as query it. So when the tool for genealogical data should not only display the data and enable queries, but I should also be able to add new individuals or relationships from the same interface (which can be network based but will also need a standalone version which will run on handheld devices and laptops for remote field work with no internet connection. [1] Sub-sets of genealogical data will vary – nuclear family, n-generations up and down from ego, n-number of collaterals from ego, spouses, males in a lineage, females in a lineage and so on – the list is endless so must be flexible for any user's needs.
Main Actors:	primary actor
Other Actors:	other actors

Services: This use case is based on a description of work being done in a joint e-science project at Durham and Kent. Progress on all these fronts and there are already individual tools that are starting to do all of the things described, but they are not yet seamlessly integrated the way we'd like them to be. Services required are as follows:

1. distributed query resulting in links to source data
2. various query interfaces
3. consolidate various data types relevant to an individual resulting from search
4. do query based on geographical information or based on individual
5. do query based on annotated video material
6. extend queries on private information to external sources
7. enter and publish data, adding new individuals and relationships
8. need standalone and networked tools for this

Contacts: Stephen Lyon (Durham)

URL: `http://bateson.dur.ac.uk`

This use case clearly illustrates the need to manage both individual research data and "shared" data in a flexble way and includes publication of data. It requires both on-line and off-line working and search facilities from a range of interfaces. The text below is from the Web site.

Indigenous Knowledge and how People apply it.

David Zeitlyn, Stephen Lyon, Michael Fischer, Paul Sillitoe.

Interactive social research methods draw together a wide range of approaches in the Social Sciences, methods that trade under names like ethnography and qualitative data analysis that are based on interaction with people where continuous analysis influences subsequent data collection, and where the research themes themselves may change during the process of research. Interactive methods have been demonstrated to be excellent knowledge building tools. Interactive methods have been primary methods in disciplines like Anthropology and sub-areas of Sociology for some time. Over the past two decades interactive methods have become more prominent, particularly in applied areas such as educational and community health research, as well as an expanding role in industry. E-Science technologies have enormous potential for advancing the quality, applicability and applications of interactive research, and directly addresses many of the greatest problems encountered when applying interactive methods. Interactive methods result in data sets that are difficult to code or otherwise organise, where the data are often difficult to compare, and where the management of data, and especially of transformations of the data such as coding, can be daunting. The access grid, storage grid and computational grid can all be leveraged to support these aspects of interactive research.

We are investigating the organisation, structure, transmission, creation and deployment of knowledge,

using interactive research methods, and exploiting middleware to support toolkits that can be adapted to specialised, research driven qualitative research tools. The objectives of the substantive research are a) to develop models of knowledge and the processes of instantiation of knowledge that improve our understanding of the dynamics of indigenous knowledge and extend our ability to describe these dynamics, b) how we can apply our understanding to policy streams relating to international development. Using existing and new data, we are addressing the study of Indigenous Knowledge, a substantial area in which research groups at Kent and Durham have an international reputation in conventional and computational approaches. We are particularly concerned with the ethical issues relating to the study and dissemination of IK. We are thus approaching e-social science research support using a largely interactive, qualitative approach. We will develop support for qualitative data and software components that address interactive collection and aggregation of data within the fieldwork segment, access to and aggregation of external data from the field site, and consolidation, analysis, modelling and dissemination from our institutional bases. This will be most visible as a distributed generalised Qualitative Data Analysis (QDA) framework that will support data collection and integration, layering, aggregation and collaborative analysis and dissemination within a grid framework and where possible within a conventional WWW environment, and in asynchronous mobile contexts, a common requirement in ethnographic research. particularly where teams are involved.

We will also be addressing the development of quantitative measures of qualitative data and analysis based on a re-orienting of Information Theory, and indeed extending our research on interactive quantitative research, trying to draw the qualitative research in as a positive means of improving quantitative research, and bolstering the role of qualitative research.

Technical Issues

The basic idea here is to adapt and integrate a raft of techniques that have been found critical (or at least very useful) in building successful information technology to support research. The present middleware project at Kent and Durham is being implemented in terms of objects we call e-data and e-documents (which is a collection of e-data, and can serve as an e-data object). E-data facilitates collaborative, complex and contingent analysis by maintaining a record of all the transactions that data is subjected to in analysis, so that at any time the derivation of a given result can be retrieved. There are a number of advantages to e-data with respect to accuracy, management, collaboration and portability. Perhaps most important, e-data can be implemented in a variety of ways as we have, using existing frameworks such as Cocoon, application servers such as JBoss, as well as more exotic forms. E-data is an object consisting of sources (e-data objects), one or more transformers and an output. Any e-data object can be a source for another e-data object. Transformers may be human where the person-work is recorded as the transformation (or indeed a reminder to someone to write an abstract!) and serves as a record of the work, such as abstracts and bibliographic references, both of which are transformations of an original information source. Transformers are also procedures that select a part of a source, reorganise a source, aggregate sources, or do some kind of computation on a source. E.g. a paragraph of a text, a concordance, a database query, a web search, a segment of a video, a video conference (not always on), a recording of a video conference. An e-document is simply a collection of E-Data, and can itself be a source for an e-data object (it is an e-data object). There are also access protocols to help address issues of privacy and ethics. E-data can help manage research as well as communicate results. E-data can represent relations between data asynchronously. The state of completion can be represented and monitored, e-data objects can defer processing until specific states are achieved, or undertake processes which facilitate progression (e.g. emailing a request to the researcher, identifying other e-data which is required by the e-data of interest) supporting a

kind of distributed critical path analysis, using a deontic logic adapted to represent ethnographic processes. References to the data can be portably moved from site to site, with everyone working on the same data, creating a kind of distributed wiki-like platform, only with greater transparency (and more security).

A.2 Archeology

Name:	Silchester Roman Town VRE
Scenario:	field workers dig finds. The photograph them in-situ to record position and contextual information. Photographs taken with a PDA are uploaded via wireless link to a field station, avoiding having to walk long distances. Experts can view the photos and identify the finds in comparison with previously discovered artefacts in a database. An example for public finds is the Portable Antiquities Database: `http://www.finds.org.uk/finds/`
Use Case:	use case
Main Actors:	field worker digging samples and recording data
Other Actors:	specialists identifying finds, and providing dating evidence
Services:	Researchers will want to do the following:

1. find new site based on surveys or other historical evidence (anecdotes)
2. identify indigenous wildlife, e.g. ducks, and proceed only if not an SSI
3. dig slowly in selected places with due care and attention
4. take photos of new extracted artefacts and grid references, use GPS
5. upload info using wireless, goes to a central finds data base
6. use software to create 3D representation of the artefact
7. create map of finds and detailed maps of finds
8. use this information to suggest other places to dig for key functions like the bath house
9. remote experts compare images of new artefacts with similar in an archaeological DB, to obtain dating information and provenance

Contacts:	contacts
URL:	`URL`

We have taken generic input from this domain by looking at the Silchester VRE project, Reading.

We also considered the LEAP project (Julian Richards, York). `http://ads.ahds.ac.uk/projects/`

leap/. This is funded by AHDS under the ICT Strategy Programme. The aim of the project is to investigate novel ways in which electronic publication over the Internet can provide broad access to research findings in the arts and humanities, and can also make underlying data available in such a way so that readers are enabled to 'drill down' seamlessly into online archives to test interpretations and develop their own conclusions.

We provide a separate use case for the OASIS project *Online AccesS to the Index of archaeological investigationS*.

A.3 OASIS

Name:	Online AccesS to the Index of archaeological investigationS
Scenario:	In England the vast majority of archaeological fieldwork is carried out by commercial organisations, which operate to specifications developed by curatorial archaeologists working in local government planning offices. Thus whilst many of the consumers of archaeological information sit within the offices and lecture theatres of universities, the majority of the producers work in the commercial or governmental sectors. Additionally the University community is rapidly losing touch with the latest developments in field archaeology as unfortunately the majority of fieldwork reports rarely enter the public or academic domains. Since 1990 an immense mountain of grey literature, approximately 17,000 unpublished reports, has grown to unmanageable proportions. Yet these data provide the primary resource for any researchers, in Britain or abroad, interested in the current state of knowledge about our heritage. See http://ads.ahds.ac.uk/catalogue/library/greylit/ The OASIS project sought to tackle the problem of a lack of knowledge about or access to the latest research data in three specific ways: 1. through the creation of a single index to the grey literature of archaeological assessment reports and excavation archives in England 2. through the provision of on-line access to that index 3. through the establishment of a mechanism to facilitate the continued collection of this research data in the long-term

Use Case:	In the past, the process of collecting information from field workers and publishing it in national monument records consisted in a number of manual tasks. In most cases the field unit undertakes their work and will produce a lot of the results in digital format which will be sent to the local and national archives (SMR and NMR) and printed out, the excavation report will then be placed in a backlog and, eventually inputted once more into a different computer. Ideally this should be handled using machine-to-machine technology reducing the human intervention and backlogs. However validation is currently handle by "experts". In local government, checks are carried out as to what the claim is, are there any relevant local monuments, parish name, field unit, etc. At the national agency there are checks on national standards, MIDAS comliance, similarity to other records, terminology, SMR, etc. This process needs to be captured in appropriate semantic and workflow services (see figure below). Most counties are now using OASIS services, but not all of them. There is shared development of standards and an inherently collabortive process with clear roles and responsibilities. Digital Curation processes are helping to reduce duplicates and provide a persistent and pervasive record. Foreseen areas of growth include addressing: Backlog and the rest of the country; Backlog bigger than front-log; Quality of grey literature; Quality of DC archaeology; Only grey literature; Geophysics? Surveys?; Only DC archaeology; ac.uk? DNA? C14? Dendro?; Closed process; Import and export issues; Only UK; Sharing data not processes; Single data source.
Main Actors:	commercial field archeologist
Other Actors:	validators and academic information consumers

Services: **Delivering OASIS to the archaeological community**

Records are created for ArchSearch, the online catalogue of the ADS. OASIS has delivered, for the first time, a fully unified record for archaeological interventions in England from around 1700 to 1998. This is made up of:

• 17,000 Concorded English Heritage Excavation Index and Archaeological Investigations Project records;

• 50,000 enhanced English Heritage Excavation Index records.

The records are catalogued according to the Dublin Core metadata element set and provide:

• the name of a project;

• a short description;

• dates of the project;

• the location of the artefactual and paper records;

• the name of the organisation responsible for the work;

• any bibliographic references;

• the geo-spatial location of the work;

• the principal types of archaeology found and their dates.

Map-based searching for ArchSearch is provided. OASIS has enabled the ADS to develop a map based search interface for ArchSearch, the online ADS catalogue of the ADS. Map-based searching is intuitive and provides easy access to the sophisticated research data held by the ADS. Such an interface is not intimidating to the novice user and quickly allows them to gain information about the archaeology of a particular geo-spatial location. A whole range of other search techniques are available for more sophisticated users, with more complex queries.

The records are also available via the AHDS Z39.50 interoperable catalogue.

Long-term sustainability for OASIS is being addressed. Archaeological excavation in England continues apace, with many thousands of excavations being carried out every year. In order to keep the scholarly community up-to-date with the latest discoveries OASIS needed to be sustainable. Consequently an on-line data capture form has been developed which will be used by contracting units and university excavation projects alike to notify the National Monuments Record of the latest archaeological activity, these records will then be passed to the ADS at six monthly intervals to be added to ArchSearch. Consequently information about the latest archaeological discoveries will be in the academic domain within a few months of the work actually taking place. A demo of the OASIS service can be seen at `http://ads.ahds.ac.uk/project/oasis/demo/`.

Contacts: William Kilbride (Head of Research in Human History, Glasgow Museum Service)

URL: `http://ads.ahds.ac.uk/project/oasis/`

Glossary

HER = Historical Environment Record (Worcestershire County Council)

NMR = National Monuments Record (English Heritage)

SMR = Sites and Monuments Record (Durham County Council)

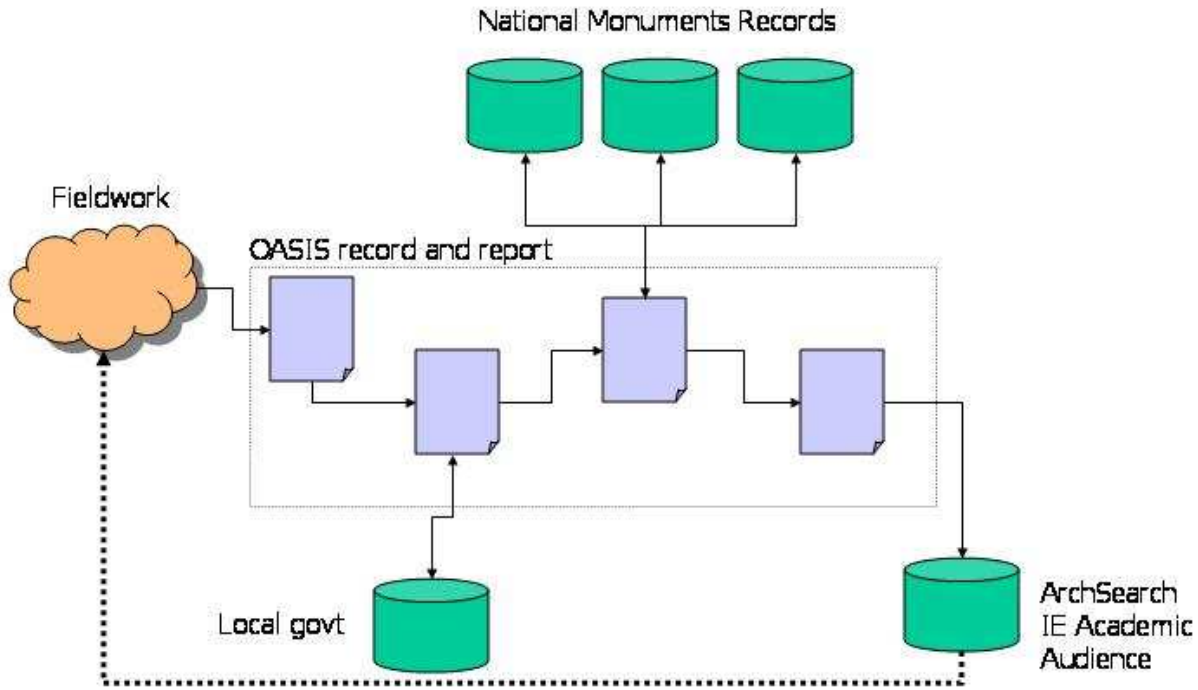

Figure 8: OASIS Architecture

A.4 GRADE

Name:	Geospatial Repository for Academic Deposit and Extraction
Scenario:	Depositing data in the geospatial data repository. Geospatial data includes any data with geographic information, such as map references, town names, post codes.
Use Case:	See below
Main Actors:	a researcher
Other Actors:	other actors
Services:	

1. Register to use GRADE repository
2. Access import/ deposit tool
3. compress files to agreed format
4. upload to submission area
5. repository system does validation checks
6. confirm copyright status
7. notification of errors to be fixed
8. create and upload metadata
9. provide quality measure
10. validate metadata and save
11. preview and re-edit if necessary
12. submit for publication
13. review by GRADE administration team
14. team moves submission to public area and sends notification

Contacts:	James Reid, Eddie Boyle and Anne Robertson (EDINA)
URL:	`http://edina.ac.uk/projects/grade/`

Full Use Case

John wishes to deposit an item of geospatial data in the GRADE repository system. He first registers to use the system, providing his personal details such as name, institution and contact details. These are stored by the repository system with due attention to the issues surrounding storage of this type of information. John then logs in to the repository system and uses the "import/ deposit" tool.

The import/ deposit tool consists of two parts, a metadata creation/ editing interface, and a data

upload interface. The first step is for John to package and compress the files comprising his data into a single file, using the common "Winzip" format. He then reads some warnings that the repository system import/ deposit tool displays about using the repository to store data, and what sort of formats are allowed. Then he uses the repository system data upload interface to transfer his zipped file to a "quarantined" area of the repository system, supplying some information about what format the data is in. There it is checked automatically by the repository system to see that it passes some integrity checks, namely: 1) that the data format is as expected, is valid, and is in an allowed format, 2) that the uncompressed data does not exceed a certain specified size, 3) that the data has some UK coverage, and 4) that the data is free from viruses. John is asked to confirm that to the best of his knowledge he is not breaking copyright restrictions by depositing the data (particularly relevant if the data is derived from other geospatial data sources) and he is also asked by the data upload interface to read and agree to repository rights information and disclaimer text.

If the data fails these checks, then John is given notification of the details of this failure, and the deposit process is stopped. If the data passes the checks, then the file is moved from the repository system quarantine area into John's personal "workspace" in the repository system. The metadata creation/ editing interface is also displayed so that John can create metadata to describe the data. Some fields will be automatically filled in already, namely John's personal details which have already been supplied by him when he registered with the repository system, and some information which was extracted from the data during the data upload checking process, namely format type, coordinate system type and extent details (in the form of coordinates). John also has the option of using a template metadata record as a starting point, or a metadata record that he created for an earlier deposit. The metadata creation/ editing interface also has a graphical map display for John to draw a box to supply coordinates instead of having to type them in (if for some reason the validation process does not produce a suitable set of coordinates), and a gazetteer interface to help with supplying placename keywords.

John is also asked to supply some measure of quality of the data; this is used by the repository to let others know what to expect if they re-use the data. He is also asked to supply rights information that he wishes to be associated with the re-use of his data.

John uses the "validate" option of the metadata creation/ editing tool to check that the metadata he has created conforms to the specification required by the repository. A mandatory number of metadata elements are required to be supplied and this validation process is a good way of checking this. Once validated, he then "saves" the metadata record in his personal workspace where it is associated with the relevant data files. At this point in the deposit process, no-one else can see the data or metadata and John can use a "preview" option to check that the data and metadata look as he wants and don't have any errors. He can re-edit the existing metadata record if he changes his mind about something or sees a mistake. He can also see any other metadata records and data that he created and deposited previously and can delete them or edit them as necessary.

When he is happy with the metadata and data, John wishes to submit it to the publicly available part of the repository. He uses the "submit for publication" function of the metadata creation/ editing tool and is told that this process will involve a review by the Grade repository administration team and is not immediate. If the metadata has successfully gone through a metadata creation/ editing tool validation check, the metadata and data are moved into a review/ pre-publication area of the repository system and looked at manually by the repository administration team; they are checking to see that there are no obvious errors in the supplied metadata and that it describes accurately what

it is in the associated data. If this check is successful, the repository administration team moves the metadata and data to the public part of the repository system, and John receives an email notification from the repository administration team informing him that the metadata and data are now in the public area of the repository system and available for all registered users to discover and download. If it is not successful, John receives an email letting him know why.

[see GRADE poster]

Plan to surface GRADE content via Go-geo! portal.

A compendium of use cases for this project was compiled by Mike Smith of EDINA [23].

A.5 R4L (Repository for the Laboratory)

Name:	Data and Metadata capture in the R4L Project
Scenario:	See below
Use Case:	See below
Main Actors:	primary actor
Other Actors:	other actors

Services:

1. devise experimental strategy subject to COSHH regulations
2. submit to system including metadata for objects and people involved
3. set permissions based on identifiers
4. recall plan on laboratory device
5. select one or more experimental techniques and appropriate workflows
6. lead laboratory worker through experimental workflow
7. automatic recording of data and metadata, including instrument and software versions
8. archive this information (data capture and deposition) for later retrieval
9. convert data formats from proprietary to common
10. submission of sample labelled with originator and identifier
11. delegation of experimental tasks
12. scheduling of experiment
13. collection of time-stamped metadata
14. assertion service
15. recording of digital images
16. assessment of quality and sustainability
17. data post processing – solution, refinement, report preparation
18. archival and curation – e.g. off-site tape store
19. search based on chemical identifier to find related information and include in report
20. selection of template for report based on experimental technique used
21. report generation in standard journal format
22. deposition of report in institutional repository
23. create links from report to underlying datasets

Contacts: Simon Coles (Southampton)

URL: `http://r4l.eprints.org`

Full Use Case

A chemistry researcher proposes a strategy for synthesising a new chemical compound and devises an experimental plan that complies with COSHH regulations. Using the SmartTea system the researcher outlines the synthetic strategy, including ratios of all reagents and solvents involved and the methodology to be used and submits it to the SmartTea system, along with metadata relating to proposed identifiers and workers involved. When the researcher is ready to commence the synthesis the plan is recalled on the laboratory tablet PC and the SmartTea system prompts the laboratory worker to measure out the required reagents and solvents in the predetermined order. The measurements are performed in the Smart lab environment, where actual values for the amounts of reagents employed are seamlessly recorded and archived for future retrieval. On completion of the reaction, once separation and purification processes have been performed, the volume or mass of product are recorded and if necessary the sample is crystallised into a solid form, the mass of which is recorded.

For both publication purposes and investigation of the properties of the new material a thorough characterisation must be performed. The research team decide to perform infra-red spectroscopy, mass spectrometry, single crystal diffraction and theoretical quantum mechanical calculations. An account for the sample is generated in the R4L laboratory data management and archival software and metadata core to all processes is generated, captured and deposited. This metadata principally comprises a (proposed) chemical identifier and the research workers involved in the study (which sets access permissions).

Infra-red spectroscopy is a technique that can easily be performed by any researcher (after brief basic training) on a desktop instrument controlled by a PC running proprietary software in a matter of minutes with virtually no post collection data correction or workup involved. The researcher prepares the sample and initiates the R4L software on the controller PC, opens the account for the sample in question and selects infra-red spectroscopy as the analytical experiment about to be undertaken. At this point, time-stamping metadata are generated and submitted to the prior assertion service. Metadata regarding the experiment are then captured and primarily include instrument (manufacturer and model) and software (including version) used and the researcher conducting the experiment. The sample is then loaded on the instrument and the proprietary software initiated. A spectrum is acquired and saved in native software format. The spectrum is then 'saved as' a file in a common exchange format (plain text / XML) and a file containing metadata on the operational parameters of the instrument during the course of the measurement is generated. The data capture service is then invoked from within the R4L software and the files pertaining to the experiment are deposited, along with the necessary metadata, in the laboratory repository.

Mass spectrometry analytical experiments are performed as a service for many researchers and the sample is to be submitted to such a facility. The mass spectrometry experiment requires a trained technician to perform decision making prior to the experiment, which may then be easily and rapidly performed with little or no post collection correction or work up of the data. A sample of the material is submitted to the service, labelled with the name of the originator and the chemical identifier. Metadata regarding the sample are also provided to enable the service to decide on the most appropriate technique for the analysis. In the R4L system the originator of the sample delegates responsibility for the mass spectrometry measurement to the service. When the sample is scheduled for measurement the service technician initiates the R4L software and selects mass spectrometry as the experiment to be undertaken. At this point, time-stamping metadata are generated and submitted to the prior assertion service. Metadata regarding the experiment are then captured and primarily include instrument (manufacturer and model) and software (including version) used and the technician conducting the experiment. The sample is then loaded on the instrument and the proprietary software initiated.

A spectrum is acquired and saved in native software format. The spectrum is then 'saved as' a file in a common exchange format (plain text / XML) and a file containing metadata on the operational parameters of the instrument during the course of the measurement is generated. The data capture service is then invoked from within the R4L software and the files pertaining to the experiment are deposited, along with the necessary metadata, in the laboratory repository.

Single crystal diffraction analysis requires decision making at numerous stages during the experiment and a lengthy data collection process, with detailed post collection data correction and work-up. The process must be performed by personnel trained in the field over a significant period of time. When the sample is scheduled for measurement the service technician initiates the R4L software and selects single crystal diffraction as the experiment to be undertaken. At this point, time-stamping metadata are generated and submitted to the prior assertion service. A suitable specimen is selected from the sample and digital images of both the sample and specimen are recorded. The specimen is then loaded on the instrument and the proprietary software initiated. Preliminary scans (binary format files) are recorded to assess the quality and suitability of the specimen. Further scans are then recorded, decisions made and parameters calculated for the scan strategy of the data collection. Data are collected, corrected, processed and reduced to a format suitable for the researcher to download and work up. The researcher downloads the reduced data onto an office PC and performs the process of working it up (solution, refinement and report preparation). When data work up is complete the data capture service is then invoked from within the R4L software and the files pertaining to the experiment are deposited, along with the necessary metadata, in the laboratory repository, whilst the raw data (binary files) are sent to an off site magnetic tape store for archival and curation.

The researcher also wishes to perform in-silico theoretical calculations to determine some of the properties of the structure of the compound. The researcher initiates the R4L software and selects theoretical calculation as the study to be undertaken and at this point, time-stamping metadata are generated and submitted to the prior assertion service. The result of the single crystal structure determination is used as the initial starting point for the study and the process of geometry optimisation and then property calculation is undertaken. When complete, the data capture service is then invoked from within the R4L software and the files pertaining to the experiment are deposited, along with the necessary metadata, in the laboratory repository.

When all investigations are complete the report generation tool is initiated and the researcher selects the identifier for the compound under study from the list of 'active' compounds in their R4L account. The researcher is presented with a list of all studies performed for this particular identifier/ compound and may select which are to be included in the report. The researcher selects, synthesis, infra-red spectroscopy, mass spectrometry, single crystal diffraction and theoretical calculations and the R4L interface presents the results for each different dataset in the study in turn. For each technique the researcher is presented with the data via an interactive interface and may select the components to be included in the final report. After processing each selected dataset for publication the R4L software automatically generates a full report in standard journal format. When the interpretations of the full study are to be submitted as a paper to a learned society journal the experimental data report is deposited in an institutional repository from the R4L software. Links to the underlying datasets in the laboratory repository are generated and enabled from the IR.

A.6 AHDS

Name:	Data Deposition into the AHDS preservation repository
Scenario:	John has recently completed a funded project and wishes to make his research available through the AHDS. See below
Use Case:	See below.
Main Actors:	John, a researcher
Other Actors:	AHDS History collection officer, other AHDS staff members
Services:	

1. researcher reads instructions
2. select an AHDS subject centre
3. decide on format for submission
4. create submission with necessary forms (meta-data)
5. submit to collection using interface
6. collection officer assigns id, completes process log, do validation etc.
7. upload to executive
8. notify submitter and iterate until complete with no errors
9. backup and preserve
10. make public copy available

Contacts:	Gareth Knight
URL:	`http://www.ahds.ac.uk/depositing/how-to-deposit.htm`

Full Use Case

John consults the AHDS web site and reads the deposit instructions located at `http://www.ahds.ac.uk/depositing/how-to-deposit.htm`.

John identifies an AHDS Centre, in this case AHDS History that closely matches the subject area of his research. He confirms the research has relevance to the subject area and is asked to post the data. If it is not, he is directed to contact another AHDS centre that has skills in the chosen field.

John consults the guidelines on acceptable deposit formats. He observes that Filemaker Pro is an acceptable format for deposit and writes it to a CD-ROM, accompanied by documentation stored as several Microsoft Word files. He also consults the AHDS Preservation handbooks (`http://ahds.ac.uk/preservation/ahds-preservation-documents.htm`) to gain a better understanding of the actions taken to preserve the content. John prints and completes the relevant data and documentation

transfer form, catalogue form, and deposit licence. He posts the CD-ROM and forms to AHDS History.

On receipt, the AHDS History collection officer assigns a unique identifier to the collection and begins a processing log for the collection. Their first activity is to authenticate that the data has been sent correctly and has not been corrupt *en route*. The collection officer attempts to read the disc and transfer the content to a staging server. An e-mail is sent to John to confirm that their data has arrived successfully. If the disc is found to be unreadable, he is invited to resubmit their data.

The collections officer produces a backup copy of the research data and begins to process it. Appropriate software is located to import the database and export the intellectual content of the database. Areas for concern are noted (e.g. the meaning of particular data is unclear and the number of rows does not match the figure quoted in the documentation) and an e-mail is sent to John, asking him to clarify the issues. At each stage, the AHDS staff member will record their action, the date on which it was performed, the time it took to perform, and any problems encountered.

To ensure continued access to the resource, the collection officer exports the FileMaker tables to a tab-delimited format that may be imported into any database software. They also export the provided documentation files to RTF and correct errors in the digital derivative. A second, distributable version of the research data is produced. The tab-delimited text files are imported into an Microsoft Access database. Electronic documentation is exported to the PDF format. The data is organized into a standard directory structure and an MD5 checksum is generated for the collection. All actions are noted in the processing log.

The collections officer creates a collection-level record for the collection, based upon the information outlined in the catalogue form submitted by the researcher.

The data is uploaded to the AHDS Executive via SCP. A backup copy is also stored at AHDS History.

The collections officer at the AHDS Executive is notified that a new transfer has taken place and validates the transfer. If the data has not transferred successfully, indicated by differences in the MD5 value, the relevant AHDS centre staff member is requested to resend the data. If the data has been transferred correctly, the Executive collections officer checks the collection for consistency. Any errors are noted and sent to the relevant staff member.

The valid collection (equivalent to an AIP in the OAIS reference model) is transferred into the preservation repository, where it undergoes regular validation and backup. A distributable copy of the collection is transferred to the dissemination server and the catalogue metadata is updated to indicate its availability.

A.7 e-HTPX

Name: An e-Science Resource for High-Throughput Protein Crystallography

Scenario: The e-HTPX project is developing an architecture and middleware services deployed as a "gateway for experimental facilities" which in the future will include the Diamond Light Source. The project comprises of:
• Focus on Synchrotron Radiation Department (BBSRC funded e-HTPX project)
• Some interest in protein crystallisation system (Oxford)
• Grid middleware, meta-data model, data collection and analysis, expert system, workflow, hub, portal
• Uses CCP4 data analysis suite (CCP = Collaborative Computational Projects, a set of UK-wide initiatives coordinated at Darsbury)
• Uses DNA expert system for image collection and control of robot plus ISpyB database for beamline data
• Hubs at SRD, OPPF Oxford, York University, ESRF BM14 Grenoble, plus services at EBI Hinxton
• Links to laboratory and commercial LIMS, e.g. through PIMS and BioXHIT projects
• Architecture for generic beamline development applicable to Diamond, ISIS, CLF, ESRF, etc.
• Industrial outreach with funding from DTI – user trials carried out at Pfizer.

Use Case: Stages in the e-HTPX workflow are as shown below. Real artefacts, such as solution well trays, crystallised protein samples, cameras, robots and X-ray beam facilities are involved in this project in addition to data management, computational analysis, metadata collection and database uploading. Perhaps the key phase for this study is the last one where meta-data and real structural information are uploaded to the Protein Data Bank. How is the PDB referenced from the IE?

Main Actors:
• Beamline staff to set up and manage experiments and robotic sample changers
• Students and professors interacting via AG

Other Actors:
• Scientists at gene expression or crystallisation facility, either a national centre like OPPF or a home lab
• Operators of shipping service
• User liaison and safety officers at synchrotron facility (SRD and Grenoble have different procedures)
• Staff running computational facilities for 3D structure determination from diffraction images
• Industrial users or providers, e.g. Pfizer
• Other people involved in PX data model, e.g. CCPn
• EBI, protein data bank admins

Services:

1. protein production – expression of gene sequences as requested by users
2. crystallisation – protein in solution in 96-well sample trays allowed to crystallise under various conditions, which may take days or months. A service is deployed which allows remote visual checking of crystals using a robot and camera. Crystals are checked daily. Image detection software is also used.
3. data collection – Prior to data collection the crystals are labelled using a bar code and categorised. A safety check at the synchrotron facility is carried out prior to shipping (crystals may be toxic or hazardous, e.g. a virus). A shipping service is used to "FedEX" the samples in a liquid-nitrogen filled Dewar and track them. They must then be maintained prior to scheduling an experiment.
4. phasing - Data collection per se consists of exposing the samples to X-ray light under strictly controlled conditions and collecting diffraction images (intensities) and corresponding meta-data. An automatic process if available as a service to do this and also controls image quality so that only as much data is collected as is required for subsequent analysis. Once intensities are available, phasing can begin. A variety of algorithms are tried, depending on the crystal type, to "interpolate" phase information. These algorithms may be tried concurrently and best results retained.
5. protein structure determination – intensities plus phases results in a set of data which can be converted into a 3D structure. This and the phasing uses computational codes from the CCP4 suite running on a dedicated cluster or on NW-GRID.
6. deposition – upload of 3D structure information and meta-data to the Protein Data Bank at EBI, Hinxton. Available for international researchers to use.

Each of these areas can (and is) sub-divided into secondary workflows and required low-level services. The data deposition step is very similar to the other use cases documented above, but concerns scientific data rather than a publication.

Contacts: R.J. Allan, D.J. Meredith and M.T. Gleaves (Daresbury)

URL: `http://www.e-htpx.ac.uk`

A.8 ePubs

Name: Using the CCLRC ePublication archive

Scenario: The CCLRC ePublication archive records the scientific output of CCLRC in the form of Journal Articles, Conference Papers, Technical Reports, ePrints, Theses and Books. It is intended for CCLRC staff and collaborators using the large-scale facilities. It is an Open Archival initiative. In the course of time ePubs will come to be a persistent and complete record of scientific activities involving CCLRC.

Use Case: A Web-based interface to an electronic publication archive. A researcher can submit an electronic copy of a publication in a variety of formats together with metadata about that publication enabling it to be found by other researchers. The submission is verified and assigned a unique identifier within the system which is recognised by the Open Archives initiative. This could initially be a preprint of a paper. If it is accepted in a peer-reviewed journal the entry could be updated to include the full publication details. ePubs can hold references, abstracts and full text, including versions of papers.

Main Actors: researcher publishing or browsing ePubs

Other Actors: ePubs developers and mainteriners. Other researchers.

Services: Non-authenticated users can browse: by year, author, affilitation, journal, report series, department, collaboration, type, title with full text. There is also a general keyword search facility.
Authenticated users can:

1. sign in
2. add new publication
3. view my publications
4. view unsaved entries
5. view draft entries
6. view entries submitted but not yet approved
7. carry out admin functions
8. sign out

A subsidiary workflow is associated with these steps, for instance adding a new publication involves completing all the information about it to create the required meta-data, uploading abstract, text, OAI, URI, DOI, etc.

Contacts: R.J. Allan and C. Jones (CCLRC e-Science Centre)

URL: `http://www.epubs.cclrc.ac.uk`

A.9 Integrative Biology

Name: Scientific Collaboration across Continents in the IB Project

Scenario:	The Integrative Biology project is bringing together an international consortium of leading bio-medical and computing researchers to address two of the most important problems in clinical medicine today: understanding what causes heart failure and how cancer tumours develop and grow. Together these diseases account for about 60% of UK deaths. Whole organ modellers (studying heart disease and cancer) are situated in 3 continents, New Zealand, UK and North America. They develop models of pharmaceutical effects on cells and how that influences the organism on a larger scale. The collaborate to develop the models and computer codes, the compare the data and visualise the large-scale effects.
Use Case:	use case description
Main Actors:	Scientists in Auckland, Oxford, Nottingham and Tulane are working together to study develop detailed, accurate multi-scale computational models of the heart and of cancer tumours. By exploiting the new Grid infrastructure to run these models on the most powerful supercomputers available for research in the UK today, they are gradually improving their understanding of these two complex systems. This will eventually lead to better control and treatment regimes. This research encompases effects ranging from DNA to whole organism modelling. In the longer term, this research will lead to an improved understanding of biological systems in general. We foresee a future where new drugs will be discovered and tested using computer models such as those which we are developing. Multi-scale models are the key to understanding the function of complex organs based on their genetic and cellular composition. The picture illustrates the different components and processes in the chain which must be modelled and integrated ranging from genetic information through cells and tissue to the behaviour of whole organs.
Other Actors:	principal scientists, developers and support staff

Services:

1. access from Web browser or Matlab
2. incorporating computable models of biological systems into executable simulation codes and installing them on a range of systems;
3. job preparation – specifying the input parameters of a simulation problem to be solved;
4. selecting the resources to be used and submitting the simulation to run
5. monitoring and controlling the simulation during execution;
6. computational steering of jobs
7. securely managing and curating both input and output datasets;
8. creating appropriate metadata information for future reference;
9. managed access to data
10. simulation subsystem which provides a range of solvers for user-supplied model codes, possibly running in coupled mode
11. storage of users files along with associated metadata
12. data retrieval
13. curation
14. visualisal examination of simulation and experimental results
15. collaborative visualisation working with remote colleagues

Contacts:	David Gavaghan (Oxford), Rob Allan (Daresbury)
URL:	`http://www.integrativebiology.ac.uk`

Requirements Analysis

Integrative Biology is typical of many large scale research projects spanning several years in that, being about innovation, its targets are hard to state specifically at the outset. The way forward emerges as the work progresses rather than being entirely predictable in advance. With this in mind, the approach taken by the project has been one of iterative development and prototyping closely involving the users. It is crucial in this type of project to engage potential users early to ensure buy-in from those who are ultimately going to use and benefit from the technology being developed.

Our initial approach to capturing requirements was to invite users to define scenarios describing how they would like to work if they had the means to do it. This was relatively unsuccessful and we eventually settled on interviewing users to try to extract as much about their requirements as they were able to articulate at that time. These interviews were based on questionnaires prepared and circulated to the users in advance. The user requirements captured in this process were sufficient to

identify the key elements of an initial prototype and the process was a good opportunity to develop the bonds between users and developers which are central to success in the project.

As the project progresses the developing prototype infrastructure is being used to implement a range of demonstrators based on individual user's research objectives. These act as a focus for critical review of the infrastructure by the users and for continuing dialogue between users and developers. In future they will also enable us to perform simple observational studies with the users.

The danger in this approach is one of providing solutions to users which they learn to live with rather than building what is really needed to make innovative progress. However flexibility and close interaction with the users should ensure that the former evolves into the latter as the project progresses. This pragmatic iterative approach to requirements capture is now showing good results in terms of both developing technical capability and growing cooperation within the project team.

Further information on this analysis and the conclusions are contained in a report of the JISC-funded IBVRE project [7].

This was centered around an analysis of the generic research cycle. The IBVRE project is tackling many of the "human centred" aspects of this life cycle, whereas the original IB project is tackling the scientific workflow which includes making shared data accessible. This is described below.

Software Architecture

An overall architecture for the project's software infrastructure has been designed to meet these requirements. See Figure 9.

The 5 main components of this architecture are:

- the user interface which runs within a normal Web browser on the user's machine as a minimum requirement or via a desktop application such as Matlab;
- a set of infrastructure components providing user-accessible services including job preparation, submission and monitoring, computational steering, managed access to data and control of visualisation;
- the simulation subsystem which provides a range of solvers for user-supplied model codes, possibly running in coupled mode;
- the data management subsystem which stores the user's data files along with associated metadata and provides facilities for data retrieval and curation; and
- the visualisation subsystem which offers the user a range of visualisation techniques for examining simulation and experimental results, possibly working collaboratively with remote colleagues.

Simulations are being carried out on a range of machines accessible to the project partners including local workstations, the commodity clusters available on the National Grid Service and the UK's high performance supercomputing facilities at HPCx and CSAR. The Storage Resource Broker, originally developed at the San Diego Supercomputing Centre, is being used to manage the wide variety of datasets which will be generated at several locations in the project. Visualisation tasks can be carried

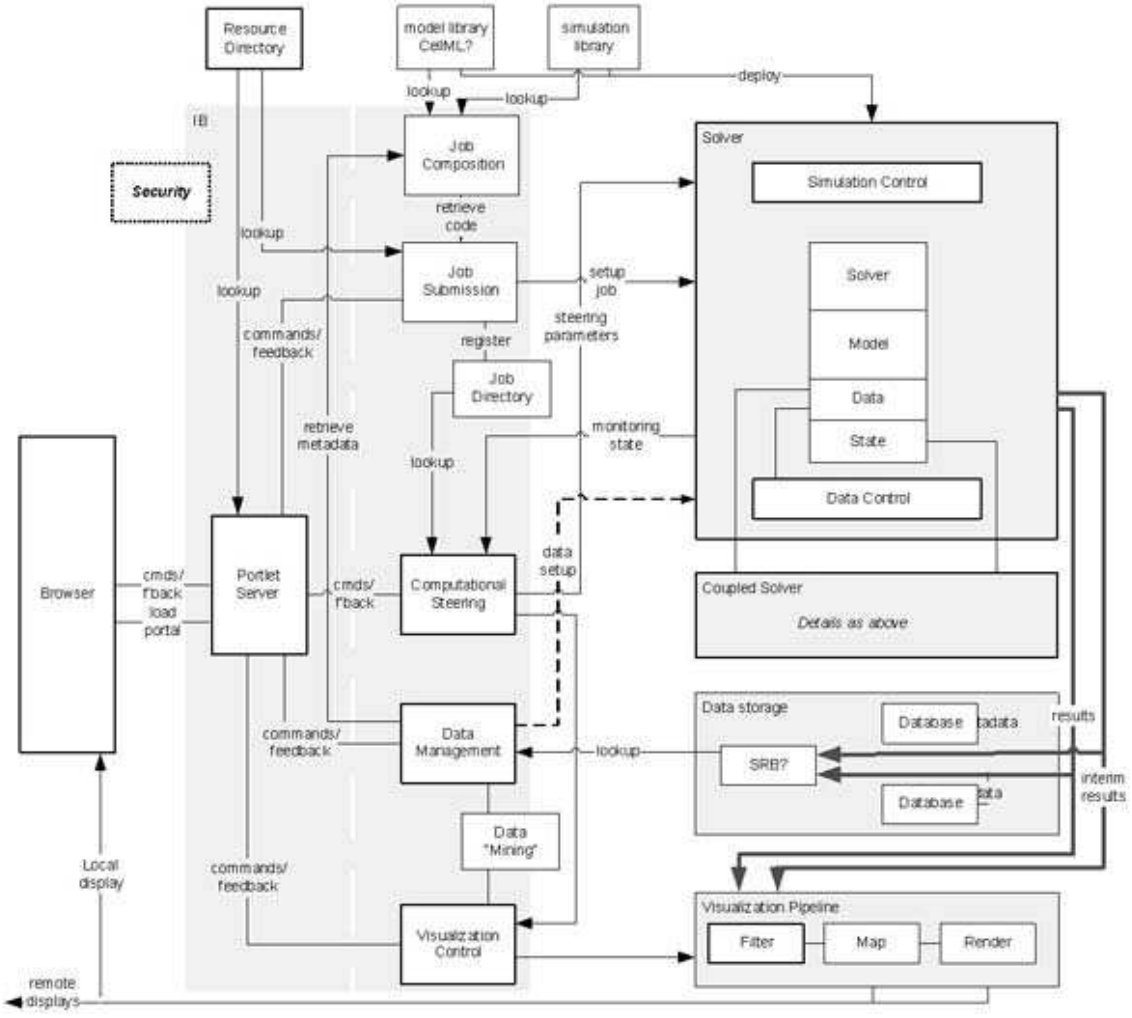

Figure 9: IB Architecture

out either locally or using Grid-based resources depending on the complexity of the visualisation functions requested by the user.

In simple terms, the main user tasks which need to be supported by the project infrastructure are:

- incorporating computable models of biological systems into executable simulation codes and installing them on a range of systems;
- specifying the input parameters of a simulation problem to be solved;
- selecting the resources to be used and running the simulation;
- monitoring and controlling the simulation during execution;
- securely managing and curating both input and output datasets;
- creating appropriate metadata information for future reference;
- analysing and visualising simulation results and experimental data;
- working collaboratively with colleagues wherever they may be.

The process of developing this infrastructure is being guided by three overarching considerations:

- working within established standards frameworks and helping to develop these where necessary;
- ensuring the software developed is scalable to address the need for increasing spatial and temporal resolution; and
- building a secure framework that will protect the integrity and confidentiality of all the project's assets.