# JISC Information Environment Portal Activity: supporting the Needs of e-Research. – **Final Report** –

**Rob Allan**
STFC 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

January 12, 2007 – updated December, 2007

**Abstract**

This is the final report of the *JISC Information Environment Portal Activity: supporting the needs of e-Research*.

The aims and objectives of this study were:

- To scope the requirements of e-Research within the area of resource discovery with reference to "portal" type services and tools;
- To identify gaps and duplication within the current provision (with reference to JISC portal and other relevant activities) therefore to identify potential areas for new work and possibly synergies that could offer a more holistic approach than currently available;
- To highlight issues and challenges that will need to be addressed in terms of serving e-Research requirements and in terms of enhancing portal activities for the IE more generally;
- To make recommendations for portal related activities that could be taken forward by JISC.

# Contents

# 1 Executive Summary

In undertaking this study we gradually came to realise that, whilst the remit was clear, the area is very broad and for a practicing scientist there are diverse requirements and dependencies, some currently outside the scope of the IE as we originally preceived it. Indeed we suggest this is one way to distinguish the Research domain from those of e-Learning and Digital Information, see [5]. For clarification, the IE is described as follows [1].

```
The JISC Information Environment is:

o an architecture that is used to plan interoperable network services
  that JISC will develop and others can provide;
o a set of standards to achieve interoperability;
o a set of projects to plan, build and contribute towards a national
  infrastructure;
o a set of services to help users access the right content from the
  right place.

It is concerned with discovery to delivery and curation of resources.
```

During this survey, we have therefore written a series of reports containing supporting information and conclusions from different perspectives. Much of this information was taken from desk-based research and by reading a large number of previous related survey reports resulting from questionnaires and focus groups targeting the research community. We have supplemented these with a small number of additional interviews, some of which were carried out with attendees of relevant workshops and conferences.

1. *Scenarios, Use Cases and Reference Models* [1]
2. *Comparison of Surveys* [2]
3. *Web-based Library and Information Services* [3]
4. *The Information Environment and e-Research Portals* [4]
5. *Interim Report* [5]
6. *A Vision for a Portal access to Global Information* [6]
7. *e-Research, Portals and Digital Repositories Workshop* [7]
8. *Final Report* [8]

This final report provides some general conclusions and collates conclusions from the supporting documents. It then lists the "top 10" requirements, as suggested to us in interviews, workshops and

[1] We thank Rachel Bruce for suggesting this text.

in reading previous surveys of relevant areas [2]. It finally gives some illustrations of the types of integration/ testbed/ development projects which could be funded to enhance the IE to meet the needs of research in a substantial number of research disciplines.

For clarity the kinds of portal commonly met with by researchers is outlined in Appendix C. This is re-produced from the Interim Report. We suggest that all the IE services should be surfaced in such portals to meet the needs of researchers.

# 2 Conclusions

We have found the key areas which need to be addressed are those of: integrating information and data; long-term archival and persistent access with appropriate access control; seamless search and discovery from a portal interface alongside other research tools; publication of data from personal and group information management systems; collaborative working in discovering, interpreting and using data and information. These broad areas, with subject-specific differences in detail and usage pattern, are constituants in the generic research life cycle and some aspects overlap with e-Learning and Digital Information management. They could thus be accommodated within the e-Framework approach, however the Joint Framework Working Group could do more to understand or address the needs of research and identify overlaps between research-oriented services and others, e.g. those of the IE. In our opinion more practical examples of how the procedures and definitions of the Framework are validated against actual research case studies would be useful. The lack of a service equivalent to CETIS (eLearning) and UKOLN (Digital Libraries) for e-Research space remains an issue when attempting to commission or develop this level of detail, however.

## 2.1 General Comments

*The meaning of the term "digital repository" is widely debated. Contemporary understanding has broadened from an initial focus on software systems to a wider and overall commitment to the stewardship of digital materials; this requires not just software and hardware, but also policies, processes, services, and people, as well as content and meta-data. Repositories must be sustainable, trusted, well-supported and well-managed in order to function properly.* (DCC Briefing Paper, 2006). See also [12].

There is a lot of work going on internationally in information management. The Information Environment is the premier UK project relevant to academic research and higher education. Some commercial research might also use the IE and its resources, and there should be an appropriate business model to encourage the use of the IE and discourage direct access of the primary sources. Likewise the use of such services as Google has to be included as a necessary part of the business model, which can promote rather than sideline the IE and deep-search services which it facilitates.

It became clear during our study that attempts to develop interoperability standards and provide single points of access to data and information are relatively immature. There are interesting issues at each stage of the process. We have taken into account end-user requirements, and could also analyse the process of publishing, discovering or accessing information. The Open Access community is starting to do this and are active in developing solutions to the various stages. This embraces self archiving

in relevant repositories (we have identified the use of institutional and facility repositories) and OA publishing (which introduces publishers' on-line repositories, e.g. domain-specific journals) [7].

In achieving many of the things we recommend, it will be necessary for the JISC to work in close co-operation with the Research Councils. In many cases the latter are already in the business of archiving scientific data and increasingly meta-data. They are also engaged in joint discussions about relevant policy. Services they provide need to be integrated with and accessible alongside IE services from user interfaces such as portals. Policies need to be compatible. Statements such as *data underpinning the published results of publicly-funded research should be made available as widely and rapidly as possible.* (RCUK, June 2005) indicate that this is a growing priority. The JISC Scholarly Communications Group undertakes work in this area and the group could offer a forum to join the IE up to the emerging policy environment.

Before giving more general conclusions and recommendations, we pull together conclusions from the various supporting study areas.

## 2.2 From the Scenarios, Use Cases and Reference Models Analysis

These conclusions arose from an analysis of a number of previously documented use cases and scenarios from relevant projects [1].

- 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 [2];
- A range of context-based user interfaces are required to access components in the extended IE architecture;

[2] this was in the language of the original e-Framework. Currently they are loosely referred to as "technical components". Reference Models are currently referred to as "Service Usage Models". See `http://www.e-framework.org`.

- 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. This includes information and data services developed in e-Science programmes in USA, Europe and Asia-Pacific regions;
- However, toolkits to support the implemtation of most of the client-side services are not yet available and portals currently provide a usable option for Web-based access.

### 2.3 From the Interviews, Surveys and Previous Questionnaires

These conclusions arose from an analysis of a number of previous surveys and questionnaires supplemented with interviews with key stakeholders [2].

Linking research practice, resource discovery and information retrieval needs an environment into which they are all integrated. We found that the previous surveys have taken too narrow a view of this, since they have mostly been discipline specific or have focussed on one aspect of this activity. The joint space requirements still need further investigation, i.e. computing and collaboration, or personal information management and admin functions. Portals tend to provide a set of customisable but pre-defined tools and services which are less flexible than a rich desktop application, although the development of interface technologies such as Ajax, and those suggested by the FLUID project are modifying this perspective in interesting directions. For an interesting example of a rich desktop application for research, see the AstroGrid Workbench `http://www2.astrogrid.org/desktop`.

However, with regard to research, some key conclusions can be drawn out from previous studies, including that:

- Researchers need access to data storage and computational resources, as well as software and services;
- Provenance is key to establishing the quality, reliability, and value of data in the discovery process (this has also been noted by the DCC);
- Any interface needs to present the views of multiple Grid services in a way that is easy for users and administrators to access and customise;
- There is a need to understand more fully disciplinary differences in user requirements (we looked more at this in [1]. Research issues which have been raised in previous studies are related to data format diversity as well as meta-data, mapping and vocabulary;
- Existing services and methodologies could be shared and Web-based presentation layers customised for delivery to users, e.g. in portals;
- A range of toolkits (thin clients, portals, scripting languages, GUIs etc.) should be developed to extend and simplify access to Grid resources and information systems leading to the eventual emergence of one or more interfaces to a Virtual Research and Information Environment. However this requires the existence of a set of underlying re-usable services. Any such services which have arisen from the e-Science Programme and JISC VRE programme are currently very domain-specific and require expert knowledge to use.

With the development of e-Research groups, new needs appear to have emerged. It is likely that the needs that will be important for a given institution will vary by the:

- Areas of research strength;
- Extent of infrastructural development;
- Strength of collaborative networks.

A portal for e-Research is likely to require the following, though we note that this list would need further expansion testing with users.

- Information and data from institutional and external sources:
  - Access to full text resources as well as tools to cross search both free and subscription based services. The results pages should identify subscription based resources and whether the user can gain access to them;
  - Access to departmental and local resources and repositories as well as external resources from one interface;
- Collaboration and research resources:
  - A mechanism to provision members (people, devices) in collaborative sessions. To include shared access to data repositories for searching, replication and updating. This requires ID and access management across institutions;
  - Applications such as Web pages, shared presentations – in an environment like Access Grid these are driven by a presenter from a master document and can also be viewed in a portal version;
  - Generic tools: text chat, white boards - need shared updates to text message streams;
  - Audio-video conferencing and collaboration tools – to share events specifying changes in compressed streams;
  - Visualisation – to share events corresponding to changes in pixels of a frame buffer, maybe using SVG;
  - Shared maps, instruments, (e.g. medical);
- Training resources:
  - Alert services, promotion and training opportunities on how to use services accessed from the portal.

Any provision also needs to be supported by assistance with organising and managing research data sets, as well as a training programme.

It is worth remembering that *a portal is not a repository, and a repository is not a portal* [10], even though all repository facilities are likely to have Web sites, these are not all portals and therefore the interfaces to their services are hard to re-use and preclude machine-to-machine access. The complementarity of portals and repositories was also noted at the workshop held as part of the project

information gathering and validation exercise [7]. Also, what is functionally useful for users is different to a portal's functionality; or, in other words, a portal could do without a repository if users are depositing in institutional and other digital repositories. In terms of e-research, the value of a portal could be enhanced it it were also an interface to create and support a community of users, a space that people use for their collaborative research. However, the portal must point at some repository or other, or it loses much of its usefulness.

It is clear from the surveys of user requirements that researchers need access to scientific and other data as well as publications. Whilst it is probably not within the remit of IE to host all such data, it may consider hosting corresponding meta-data or providing search facilities and mechanisms to link data to publications and *vice versa*.

Overall, researchers appear to need more support for learning, adapting, and writing software specific to their research problems than is currently available. Also, researchers who are generating and using large data sets need help managing their data. This need will become more pressing as data enters long-lived data repositories and therefore the public arena through preservation rather than simple publication of links.

We suggest that the IE might usefully link into a wider range of data services since it has developed a number of re-usable components within the Technical Architecture. These include private and commercial as well as open services.

As an example, which illustrates the complexity of requirements, we cite a recent study performed to select datasets of primary interest in the ESRC e-Infrastrucutre project [14]. This mentioned the following datasets:

- British Houshold Panel Survey
- Census 1991 SARs (Samples of Anonymised Records)
- EDINA UKBorders Census boundary data for SARs 1991
- Health Survey for England
- National Child Development Survey
- Datasets from MRC and NERC as required
- Quarterly Labour Force Survey
- General Houshold Survey
- British Social Attitudes Survey
- Workplace Employee Relations Survey
- ONS Omnibus Survey
- ONS Millenium Cohort Survey
- International Monetary Fund (IMF)
- World Bank and Organisation for Economic Cooperation and Development (OECD)
- ESDS International service
- British Crime Survey
- ONS Neighbourhood Statistics
- European Social Survey
- Eurobarometer data series
- Administrative, retail, consumer, video CCTV and Web usage data

These datasets are from diverse sources, local, national, international; open, commercial and confidential. The same survey mentioned the following requirements related to access and tools.

• Shibboleth enabling
• GIS system to utilise boundary data
• Longer term access to data via Grid technology direct from provider
• Software and methodology from existing projects
• Access to datasets from other disciplines
• Tools for geographic mappings
• Meta-Data registries
• Controlled vocabularis and ontologies
• Question banks
• Classification schema and variable mappings
• Linking between data and related publications
• Make commercial tools such as SPSS, SAS, Stata available for Grid work
• Virtual safe setting for analysis of confidential data, e.g. Census Controlled Access Microdata Samples.

Whilst this is illustrative of the broad field of social science, similar data requirements and diversity of sources is typical of many research fields.

# 3 Top 10 Conclusions and Comments on our Vision

At the workshop held as part of the project information gathering and validation exercise [7], we explained a vision of the portal as a "one-stop-shop" for researchers to access data and global information relevant to their studies with appropriate semantic support and contextual services for discovery, location and digital rights management.

Putting the outcome of the workshop together with the outcome of other surveys which we analysed [2], we identified the following 10 conclusions.

1. IE needs to work alongside Google and other popular search engines like Yahoo!, Lycos, MSN and Ask which provide "generic search" functionality. Meta-data needs to be provided such that harvesters provisioning these search engines can locate appropriate UK HE and e-Research services providing "deep search" functionality;

2. IE needs to provide mechanisms for discovering scientific data and linking between data, publications and citations. This could potentially be based on methodology developed in the Claddier and SToRE projects;

3. IE needs to use protocols and standards to facilitate exposing its services in a variety of user interfaces, including portals;

4. IE needs to provide facilities to publish from and to make available content from personal and group information services. This will be of particular significance if Collaborative Research Environments or other systems supporting distributed communities are to become embedded in institutional practice;

5. IE should embrace such things as JISCmail archives which contain research material (although this needs its original context);

6. Training for users is needed as well as increased awareness of what is available. It was stated that being able to seek advice from an "information professional" on deep searches or alternative/ related sources of information was useful;

7. IE needs to embrace Semantic Web technologies, including provenance. This could include adopting domain-based ontologies to broaden search terms to similar but not identical areas;

8. IE needs to work with and provide enhanced access for commercial sources and inter-operate with proprietary software. This will avoid isolation of potentially important resources;

9. IE interfaces (e.g. portal) need to be highly customisable and to treat users' searches "in context";

10. IE should embrace collaboration technologies to facilitate joint uses of its services (multiple users and multiple services interacting). This is not simply about controlling access, but includes how to carry out steps in the research information lifecycle which require on-line collaboration, e.g. joint editing of reports using a Wiki.

# 4 Recommendations to the JISC

Perhaps in conjunction with the Research Councils, JISC needs to articulate clearly the extent to which it supports e-Research. There are opportunities to work with the Research Councils and build on synergies, but this requires discussion at a higher level and a carefully coordinated programme of work. There are other non-RC providers and consumers – some which spring to mind are: the BBC, the Ordnance Survey, the Met Office – many researchers depend on these too [3].

Some recommendations for such a programme, based on our surveys and interviews are listed below.

1. Research Requirements "from discovery to delivery":
   (a) Address integration of data and information
   (b) Need to include data (and associated meta-data) from research areas
   (c) Researchers (at least some) want to work collaboratively, so we need portal-based tools for shared interpretation, shared authoring/ publication, etc.
   (d) We have defined a "research life cycle" which reflects the above scenarios, see Appendix B. We have also collected a number of supporting use cases and analysed what high-level services these imply.
   (e) The life-cycle could be extended to include sharing and publication of information and data which was private to an individual researcher, perhaps using P2P technology: private information → sharing with peer group → publication
   (f) Implications are that new "business models" are needed (for both academic and commercial publishers). Also need a new "review model" to moderate making information and data public subject to appropriate validation. This would include incentives for self archiving and linking publications to data and ongoing mechanisms for adding citations to meta-data

[3] We have not considered issues of IPR here, this is not our area of expertise but we recommend that JISC should investigate this.

(g) questions of ownership have to be addressed for data use and re-use. This implies extensions to current mechanisms for digital rights management

(h) Awareness raising and training is also needed, as was noted in several past surveys.

2. Portlets for the IE, some practical suggestions:

(a) Portlets are "portable" and can be embedded in portal frameworks: institutional; service; project.

(b) Portlets are a "veneer" to underlying services. This implies a rich set of robust and highly-available IE services are required. APIs, access mechanisms "T-Models in the terminology of UDDI", and semantic descriptions must be provided in registries so that the services can be located and used.

(c) Portlets are customizable:

- Need a "reference set" and "repository"
- Need HCI study or engage CSCW researchers to help understand which functionalities work best together

(d) Take best current solutions, e.g. HEIRPORT, CREE, SPP, and extend these to include new tools. See examples below.

(e) One size does not fit all. Also need data/ information accessible on the desktop

3. IE Architecture, comments on technical issues:

(a) We had some difficulty in understanding what has been delivered by projects to date, what is robust, what is re-usable, where to get the software?

(b) Implies a need to produce a "map" of services which fit above the architecture design, and also a need to pro-actively maintain them

(c) Do they "fit" the e-Framework methodology?

(d) Need to satisfy requirements of (1):

- Bring in services for data/ meta-data
- Seamless Search/ discovery
- Fine-grained Access control
- Packaging/ access/ analysis
- Collaboration
- Links to computational services such as NGS
- Links to other services and software developed in RC-funded projects (not only e-Science)

(e) New protocols need to be included in addition to Z39.50: OGSA-DAI, WS-RF, etc. to enable IE services to inter-operate with Grid-based research services.

(f) New core services will be needed: e.g. a "super registry" combining IESR and UDDI; or a "super aggregator"; or a "citation maintainer" (including data linked to publications); or a "data resolver".

(g) Need semantic descriptions of services for registries. Also semantic descriptions of data based on available meta-data

- Subject specific vocabularies

- Implied need for ontologies to map between subjects and widen search criteria

(h) Take experience from projects such as DCC and scientific data archives such as BADC into account and adapt any tools/ services which they already have

It is hard to address all these items, so we suggest a number of examples which might each address a subset. These examples are meant to be illustrative only and are not suggestions of actual projects to be funded (althought they could be). They are based on our findings of some of the more mature relevant services and scenarios identified.

Each example is specific in nature, but illustrates ideas that could extend to a different subject area and include collaboration tools in later work. We recommend the aim to make the outcomes of such projects generic and re-usable as far as possible. We considered to put timescales on the development of these services, but that would be rather speculative and probably therefore not very useful.

### 4.1 Example 1: Geospatial Data and Information.

This example illustrates work which could be composed from services already being developed or put into production in a number of projects: GRADE; geoXwalk; HEIRPORT; Go-geo!; DCC; and projects delivering things such as OpenURL resolvers and aggregators. Work on geospatial data is already relatively advanced as several JISC-funded projects at EDINA and elsewhere have addressed this area. There are also a number of RC-funded projects, for instance those supported by ESRC and NERC. See also DCC Briefing Paper *Curating Geospatial Data* (August 2006).

Figure 1 shows how Go-geo! links into the JISC landscape (from James Reid, EDINA). Note the reference to the VRE Programme which is where portal intrfaces would feature.

An integration project might enhance existing IE components to include geospatial data plus census data and associated publications. Some IE projects are already investigating this, for instance GEMS aims to make such data available via data-management services hosted on the NGS. HEIRNET is using geospatial data to enhance research into historical monuments and finds. Work on linking geospatial data to other relevant data and publications still needs to be increased in other disciplines where studies are made of historical and current activities, particularly in social science.

Other projects which might usefully contribute or benefit from this include: Silchester Roman Town VRE; OASIS.

In the case of map-based data, there is a need to embrace the activities of the Ordnance Survey. Whilst not completely free to use, the Survey's map information is widely used, and often forms the basis of valuable "derived data" some of which may relate to a particular time or event. Sharing such data can maximise research (increasingly multi-disciplinary) and/ or business benefits [13].

A branch of Natural Language Processing is automatic place-name disambiguation. There are a few projects active in geospatial analysis, such as cpGeo (University of Sheffield). To a great extent these are still research projects. Guided dis-ambiguation, perhaps using gazetteers (authoritative lists) such as fuzzyG (Hof University, Germany) provides similar functionality. This is relevant in the case of name-based geospatial data.

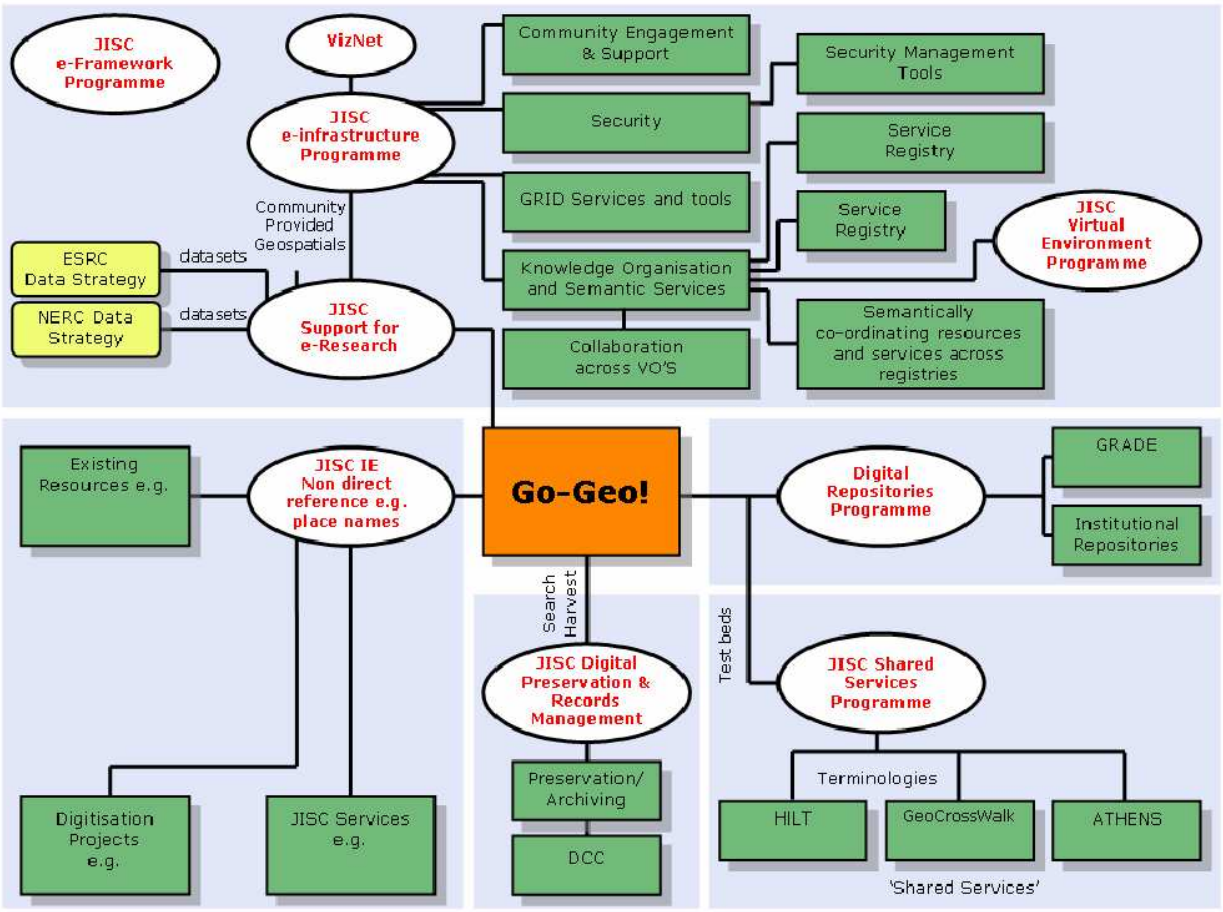


Figure 1: Go-geo! in the JISC Landscape

To facilitate multi-disciplinary work and cross searching, appropriate meta-data should be used, not just that specific to geospatial data.

The DCC is currently working with EU partners in proposing an ISO standard for digital archives, e.g. relating to the the level of service provided which will include a guarantee of the longevity of the data. All data and information holdings which are expected to be persistent should be certified to this standard as proposed by the DCC.

Some required services include:

- user registration
- authentication
- meta-data creation/ editing
- packaging
- data/ information upload
- validation
- notification
- rights information creation/ editing
- authorisation
- recording/ editing provenance
- representation information repository
- search/ discovery (possibly in collaboration)
- locate (based on rights)
- aggregation/ overlay
- visualisation or analysis (possibly in collaboration)
- access/ download
- add links: proposals/ data/ publications/ citations
- resolve and interpret terms (meta-data and "names")

### 4.2 Example 2: Chemical Data and Information.

This example is based on the eBank, e-Prints, PSIGate, ALPSP, eCrystals, SMART-Tea and R4L projects. Other projects such as e-CCP are related.

Figure 2 shows the eBank services exposed via the PSIGate portal (an SPP portal for RDN). The eBank research life cycle was shown in [1]. They had a similar life cycle for education.

Other projects which might usefully contribute or benefit from this include: Integrative Biology VRE and e-HTPX. Computationally generated data produced by the UK CCP Programme (the Collaborative Computational Projects) might also be included – for instance CCP1, CCP4 and CCP5 are now generating meta-data and storing results as enabled in the NERC-funded e-Minerals and BBSRC-funded e-HTPX projects. Protein structure data produced by e-HTPX and its associated meta-data are uploaded, validated and archived in the Protein Data Bank (PDB) at the EBI, Hinxton.

It is necessary to capture data and meta-data as they are generated in the laboratory, often in "one-shot" or high throughput experiments. Once such data is captured in a laboratory information management system or electronic lab notebook, it can be managed and used in other parts of the research


Figure 2: eBank and PSIGate Portal

life cycle, for instance the automatic generation of laboratory reports.

There is a similar set of services required to those in Example 1. However in this case we are handling chemical and bio-chemical data rather than geographical data. Place names and map references are replaced by chemical names, compounds and identifiers such as InChI, LSID and terms from schema such as CML. Semantic services need to be provided to work with such identifiers and terms. There is a further similarity to the place names, in that chemical and bio-chemical names and identifiers are not unique and have changed with time.

In the e-CCP project, application-specific vocabularies have been defined and semantic tools based on ontologies are used to map data between applications. Thus output from one application can be visualised with tools designed for a different one. So far this has addressed the computational chemistry domain and is based on work on the XML-based Chemical Markup Language, CML. Extensions of the generic e-CCP methodology (now known as AgentX) are however applicable to other areas and there is a requirement to store the vocabularies and ontologies in a registry such as IEMSR for re-use.

### 4.3 Example 3: Social Science Information and Data.

ESRC has recently established a small project to deliver a prototype e-Infrastructure for Social Scientists. This project makes use of collaboration tools and middleware being developed in JISC VRE projects such as GROWL and the Sakai Demonstrator to access NGS resources and data sources such as UKDA and MIMAS. It plans to link to Census data via OGSA-DAI middleware and services being deployed in the GEMS project by extending the GROWL tookit. Geospatial data is also involved, for instance in defining land boundaries for policy makers. The outcomes of the project, which will also deliver a range of semantic tools and possibly use AgentX, could be incorporated into existing IE services such as the IESR or IEMSR. In Section 2.3 we noted the complexity of requirements in this project as documented by Miller [14].

This project does not however currently have any provision for using other IE services, such as other means of data and information discovery or publication of results – these are areas which could usefully be explored. A simple first step would be to include the SPP portlets in the Sakai portal to be deployed at NCeSS hub and nodes, thus enabling interaction with SOSIG and other RDN (now Intute) and institutional repositories. Portlet interfaces to MIMAS and UKDA and other services useful to social scientists should be explored as should licensing of widely-used commercial software to run on the NGS and linking of remote datasets into such software. These services should offer an API (such as Web services) by which they could "plug in" to whatever portals or rich desktop client applications projects provide for their researchers.

Particularly in research fields like Social Science, personalisation and customisation is important. It helps the researcher to formulate questions and use the data to support their hypotheses. Data is subject to interpretation, the results of which need to be pushed into an information environment where they can be shared with colleagues, possibly providing alternative insight. An example would be an economist working with a child psychologist to interpret the effect of changes in school policy on educational attainment.

For social scientists investigating worldwide trends, other sources of data and information outside the UK will also be required.

Another worked example scenario of a social science researcher was given in [1].

### 4.4 Example 4: Email, Data and Shared Services.

This example is based on JISCmail which, as well as providing a list-server and management tools, has a searchable archive, see `http://www.jiscmail.ac.uk`. JISCmail is an example of a tool which already supports thousands of "virtual communities", each via its own mailing list. Community membership overlaps, but JISCmail maintains separate content.

DCC Briefing Paper *Curating Emails* (April 2006) notes that e-mail is now a significant tool in decision making for both administrative and research purposes. Records of how such decisions are made are no longer documented separately, and curating e-mails and attachments may be the only way to provide an audit trail in the event of problems or for future reference. For such an archive to be widely useful, search and discovery mechanisms need to be provided alongside the tools to mark up e-mail in a useful way (e.g. separating out the "to" and "from" headers as XML elements). This is a challenging area as many e-mails typically contain abbreviations, colloquialisms and material which is simply not directly relevant. There is often no reference to formal "threads" of conversation and the "subject" headers may be wildly misleading. This area requires as much a culture change as software services and could be a fruitful area for longer-term work.

Portal interfaces for search and discovery of e-mail archives, with appropriate rights-based mechanisms for access may become increasingly useful, but only if supported by semantic and text mining tools. Access rights may usefully be more fine grained than at the list level, but there is no mechanism for capturing this granularity or other meta-data not already in the mail headers. Sakai is a portal which also supports virtual organisations through its "worksite" structure. Each worksite could therefore easily be linked to a different JISCmail list. Additional worksite services would then be immediately available to JISCmail users.

The utility of such services is already indicated by the fact the Google searches often turn up e-mail contributions to thread-based discussion fora, particularly for resolving problems with computer software. Such results however also confirm the potential problems outlined above. Google also cannot access e-mail archives which are stored in a database. These should be linked into the IE in other ways enabling cross searching.

We note that JISCmail is considering additional services related to collaboration, for instance a calendaring service to support setting up meetings. This is already available in Sakai, but would require further interation. We believe that additional integration is also necessary with institutional services, such that JISCmail and calendaring can be accessible closely-coupled with MS Exchange.

### 4.5 Example 5: Large-scale Experimental Facilities Information and Data.

The new Diamond Light Source experimental facility on the Harwell Campus in Oxfordshire represents the largest single investment in science in the UK for 30 years, see `http://www.diamond.ac.uk`.

During an exercise to document requirements and develop a workplan for the roll-out of e-Science technology to support DLS, we identified the need to preserve experimental data and associated meta-

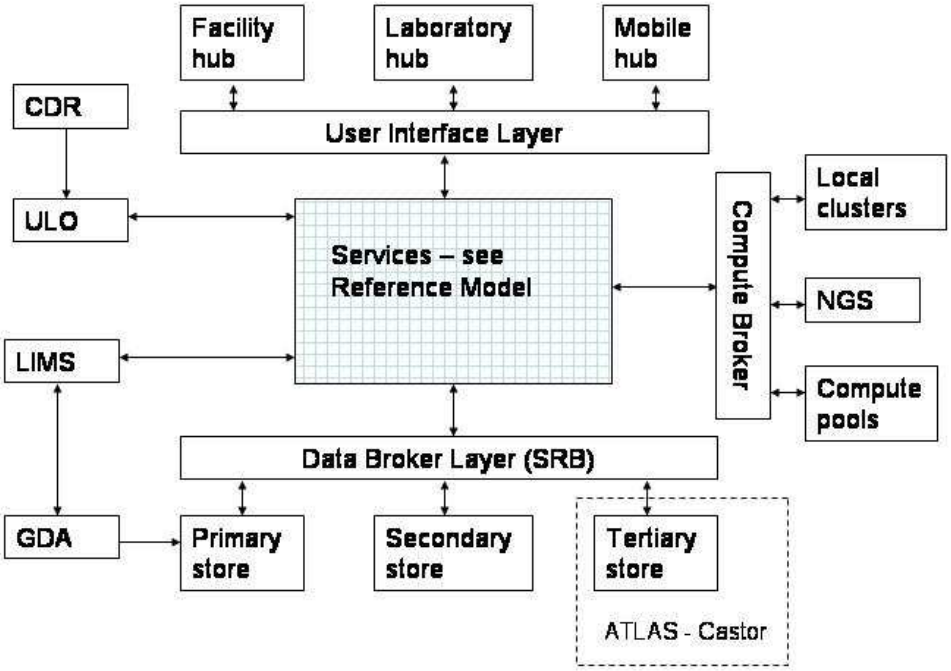


Figure 3: Diamond Light Source e-Infrastructure

data and link it to both grant proposals made to the Research Councils and publications arising from the work [11].

Figure 3 shows the general architecture being adopted for the Diamond Light Source e-Infrastructure. Some components of the proposed system include: laboratory information and control systems (PIMS, iSPyB, GDA); data management and storage (SRB, Castor, ATLAS Data Store); ePubs repository; ESRF proposal system. DLS users will also access the NGS to run data analysis and simulation jobs.

At the Workshop [7] there was a discussion about the difference between subject-specific or institutional repositories and where researchers would deposit their publications. It was considered that they would be unlikely to want to use a repository that belonged to a facility like Diamond, because it fits neither of the above categories. However the facility managers will want this, or at least to be able to easily access all publications and reports produced by their user community, as would any other service. This is important in agruments for sustained funding and IE could play a valuable role in this.

A number of possible early users of DLS were consulted in elucidating requirements, including ones from e-HTPX and the National Crystallography Service who are also using eCrystal, eBank, R4L and PDB as noted in Example 2. However, a considerable amount of work still has to be done to deploy all the required services and link them into other services, such as those provided by NGS, DCC, CCPs and ePubs. It would be helpful if JISC IE services could be used for this work. A portal framework integrating interfaces to the various administrative, training and research services from the views of users, user office staff and station scientists still needs to be specified and developed.

# 5 Acknowledgments

JISC for funding.

We thank people we spoke to: Ann Borda, Maia Dimitrova, Matthew Dovey, Bill Olivier, Andy Powell, Ian Dolphin, Debbie Franks, Roddy MacLeod, Catherine Jones, Chris Awre, Robert Sherratt, Simon Hodson, Graham Pryor, Derek Sergeant, Simon Coles, David Giaretta, Chris Higson and James Reid, Penny Windebank,

# References

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

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

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

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

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

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

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

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

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

[10] P. Burnhill (EDINA) (personal observation, May 2006)

[11] M.T. Gleaves, R.J. Allan, A. Ashtun, et al. *User Requirements and Project Documentation for the Diamond e-Infrastructure* (CCLRC, May-September 2006)

[12] P. Lord and A. MacDonald *JISC e-Science Curation Report* (JISC, 2003)

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

[14] K. Miller *Primary Selection of Datasets* Deliverable D1.1.1 of the ESRC e-Infrastructure Project (August 2007) `http://www.ncess.ac.uk/services/research`

# A Glossary, Abbreviations and URLs.

A glossary with many relevant entries can be found at: http://www.grids.ac.uk/ReDRESS/glossary_v2/glossary_v2.html.

Wikipedia can be used to obtain an explanation for most of the generic ones, http://www.wikipedia.org.

Specific abbreviations used in this report are:

**Access Grid:** virtual interaction technology supported by JISC, see http://www.agsc.ja.net

**ADS:** Atlas Data Store - petabyte tape store at the Harwell Campus

**AstroGrid:** UK virtual observatory project funded by PPARC, see http://www2.astrogrid.org

**BADC:** British Atmospheric Data Centre, one of the NERC-funded UK data centres, see http://www.nerc.ac.uk/research/sites/data/

**Castor:** CERN Advanced Storage Manager http://castor.web.cern.ch/castor/

**CCLRC:** Council for the Central Laboratory of the Research Councils, now part of STFC

**CCP:** Collaborative Computational Projects http://www.ccp.ac.uk

**CML:** Chemical Markup Language, see Wikipedia

**CREE:** Contextual Resource Evaluation Environment, JISC-funded project http://www.hull.ac.uk/cree

**CSCW:** Computer Suppoerted Collaborative Working, see Wikipedia

**DCC:** Digital Curation Centre http://www.dcc.ac.uk

**Diamond:** Diamond Light Source, see DLS

**DLS:** Diamond Light Source http://www.diamond.ac.uk

**DRM:** Digital Rights Management, see Wikipedia

**e-CCP:** e-CCP project on data interoperability http://www.grids.ac.uk/twiki/bin/view/ECCP/WebHome

**e-HTPX:** An e-Science Resource for High-Throughput Protein Crystalography, BBSRC/ DTI funded project, http://www.e-htpx.ac.uk

**ePubs:** STFC ePubs open access repository http://epubs.cclrc.ac.uk

**ESRC:** Economic and Social Research Council http://www.esrc.ac.uk

**ESRF:** European Synchrotron Radiation Facility http://www.esrf.eu

**GDA:** Generic Data Acquisition system http://www.gda.ac.uk

**GEMS:** Grid Enabled MIMAS Service, JISC-funded project http://pascal.mvc.mcc.ac.uk:9080/gems

**GROWL:** Grid Resources on Workstation Library, JISC-funded VRE-1 project http://www.growl.org.uk

**HCI:** Human Computer Interface, see Wikipedia

**HEIRPORT:** http://

**IEMSR:** Information Environment Meta-Data Schema Registry http://iemsr.ac.uk

**IESR:** Information Environment Schema Registry http://iesr.ac.uk

**InCHI:** International Chemical Identifier, see Wikipedia

**iSPyB:** Information system for protein crystallography beamlines http://www.esrf.eu/UsersAndScience/Experiments/MX/How_to_use_our_beamlines/ISPYB

**JISCmail:** JISC mail service http://www.jiscmail.ac.uk

**LSID:** Life Sciences Identifier, see Wikipedia

**MRC:** Medical Research Council http://www.mrc.ac.uk

**NCeSS:** National Centre for e-Social Science http://www.ncess.ac.uk

**NERC:** Natural Environment Research Council http://www.nerc.ac.uk

**NGS:** National Grid Service http://www.ngs.ac.uk

**OGSA-DAI:** Data Access and Integration using Open Grid Services http://www.ogsadai.org.uk

**ONS:** Office for National Statistics http://www.statistics.gov.uk

**P2P:** Peer-to-peer, see Wikipedia

**PDB:** Protein Data Bank, a service hosted at EBI, Hinxton http://www.ebi.ac.uk/msd/

**PIMS:** Protein Information Management System http://www.mole.ac.uk/lims/project/ funded by BBSRC

**PPARC:** Particle Physics and Astronomy Research Council, how part of STFC

**RCUK:** Research Councils UK http://www.rcuk.ac.uk

**RRM:** Research Referemce Models, now referred to as SUMs in the e-Framework, http://www.grids.ac.uk/Papers/Classes/classes.html

**Sakai:** Sakai collaborative learning framework adapted for research purposes http://www.grids.ac.uk/Sakai

**SPP:** Subject Portal Project http://www.portal.ac.uk/spp

**SRB:** Storage Resource Broker http://www.npaci.edu/DICE/SRB/

**STFC:** Science and Technology Facilities Council, an amalgamation of CCLRC and PPARC, see `http://www.stfc.ac.uk`

**SUM:** Service Usage Models as defined in the e-Framework for Education and Research, see `http://www.e-framework.org`

**SVG:** Scalable Vector Graphics, see Wikipedia

**UDDI:** Universal Description, Discovery and Integration, a Web services registry specification, see Wikipedia

**UKDA:** UK Data Archive `http://www.data-archive.ac.uk`

# B Research Lifecycle

We consider the activities involved in doing research to be ultimately driven by knowledge creation. We make the following definitions (as presented to members of CURL in a meeting in October 2004):

**Data:** bits and bytes arising from an observation (non-repeatable), an experiment (repeatable) or a calculation;

**Information:** relationship between items of data of the form "A is always associated with B in some way".

**Knowledge:** understanding of causality in relationships "B happens after A because of X". This knowledge is shared globally.

In Figure 4 we show our own version of the research lifecycle steps which we believe to be appropriate to extended IE activities.

We have omitted from this the activities involved with grant proposal and funding, admin, collaboration forming, actual collaboration and computing as these are probably outwith the forseeable IE activities, but may be appropriate to other JISC areas.

# C What kinds of Portals will be met by Researchers?

The researcher is likely to meet Web browser-based portal technology in three situations: (1) the Institutional Portal provided as a gateway to the services and information of an institution or large facility and maintained by central IT staff; (2) a Project Portal with all the resources of a particular multi-institution research project – a Virtual Organisation – probably maintained by project staff part time; and (3) a Service (subject-specific) Portal provided for access to a specific service, e.g. a national data center, maintained by payed IT staff as part of the service.

The following definition is from Wikipedia `http://www.wikipedia.org`: *Web portals are sites on the World Wide Web that typically provide personalized capabilities to their visitors. They are designed*

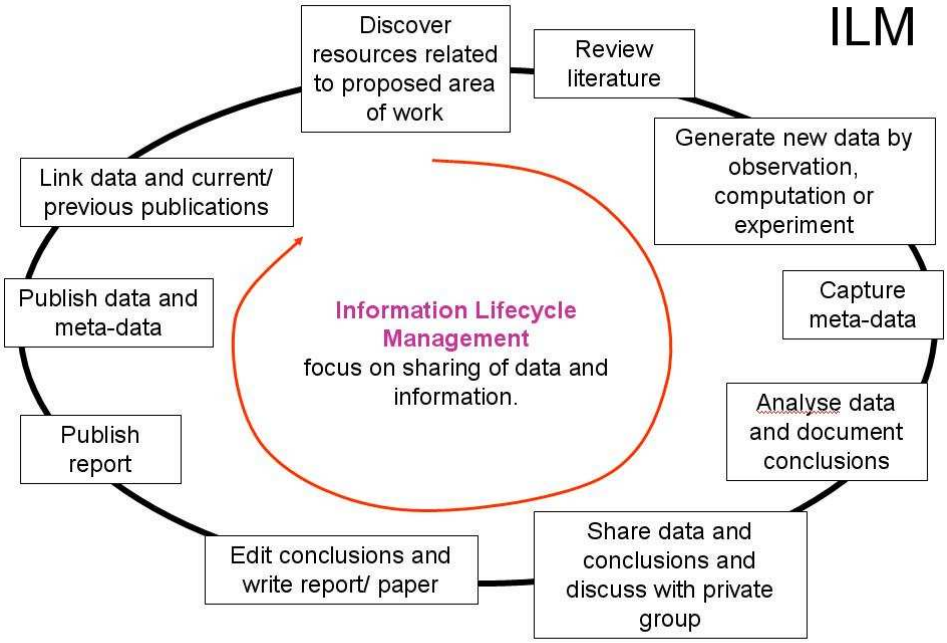


Figure 4: e-Research Data and Information Lifecycle Management

*to use distributed applications, different numbers and types of middleware, and hardware to provide services from a number of different sources. In addition, business portals are designed to share collaboration in workplaces. A further business-driven requirement of portals is that the content be able to work on multiple platforms such as personal computers, personal digital assistants (PDAs), and cell phones.*

*Many of the portals started initially as either Internet directories (notably Yahoo!) and/ or search engines (Excite, Lycos, AltaVista, infoseek, and Hotbot among the old ones). The expansion of service provision occurred as a strategy to secure the user-base and lengthen the time a user stays on the portal. Services which require user registration such as free email, customization features, and chatrooms were considered to enhance repeat use of the portal. Game, chat, email, news, and other services also tend to make users stay longer, thereby increasing the advertisement revenue.*

Different types of portal are defined to include: Regional Web Portal; Government Web Portal; Enterprise Web Portal.

## C.1 Institutional or Facility Portals

Wikipedia goes on to say: *In the early 2000s, a major industry shift in Web portal focus has been the corporate intranet portal, or "enterprise Web". Where expecting millions of unaffiliated users to return to a public Web portal has been something of a mediocre financial success, using a private Web portal to unite the Web communications and thinking inside a large corporation has begun to be seen by many as both a labor-saving and a money-saving technology. Some analysts have predicted that corporate*

*intranet Web portal spending will be one of the top five areas for growth in the Internet technologies sector during the first decade of the 21st century.* We might also refer to these as "Institutional Portal". They could be designed for or provide views for a variety of purposes: e-Learning, e-Research, Information Management, Administration, etc.

In this context Gartner defines "higher education" portals as *enterprise portals integrated with administrative, academic and other applications of interest to students, faculty and staff.* They place them high up on the "slope of enlightenment" in their 2005 HE hype cycle because, although budgetary constraints have slowed down adoption, they are emerging as key institutional interfaces for online resources and applications.

Many universities have started to develop portals, usually starting with a student portal and then moving onto other stakeholder groups, e.g. prospective students, staff, alumni. These can use portal software, e.g. Luminis, or can utilise the portal features of other enterprise software, e.g. Oracle or WebCT. Open source portals are in development, e.g. uPortal. Other organizations such as Research Councils are developing their own portals (e.g. ESRC Society Today, `http://www.esrcsocietytoday.ac.uk/`). STFC is investigating portals for access to large-scale experimental and computational facilities.

There were two institutional research portal projects being piloted under the JISC VRE programme. ELVI (Evaluation of a Large VRE Implementation) at Nottingham University `http://www.nottingham.ac.uk/research-systems`, and EVIE (Embedding a VRE in an Institutional Environment) at Leeds University `http://leeds.ac.uk/evie`. These sought to evaluate the embedding of research tools into institutional portals.

Some features of enterprise portals are:

- Single point of contact – the portal becomes the delivery mechanism for all business information services (one stop shop);
- Collaboration – portal (institution) members can communicate synchronously (through chat, or messaging) or asynchronously through threaded discussion and e-mail digests (forums) and blogs;
- Content and document management – services that support the full life cycle of document creation and provides mechanisms for authoring, approval, version control, scheduled publishing, indexing and searching;
- Personalization – the ability for portal members to subscribe to specific types of content and services. Users can customize the look and feel of their environment;
- Integration – the connection of functions and data from multiple systems into new components/ portlets.

Most enterprise portals provide single sign-on capabilities to their users. This requires a user to authenticate only once. Access control lists manage the mapping between portal content and services over the portal user base. This is facilitated by a Corporate Data Repository within the institution.

## C.2 Project Portals (Science Gateways)

Whilst an Enterprise Portal might be very good for e-Learning and Administration, as shown in the Lumenis demo, they provide an outward-facing representation of the processes and community within a single institution or organisation.

A Project/ Grid Portal used for e-Research will typically be used by people from many organisations. We will refer to this grouping of people and underlying resources as a "Virtual Organisation".

The logic underlying a Project Portal must facilitate sharing of data and resources within the Virtual Organisation which means across institutional administrative boundaries. Typically this requires Grid Middleware to comply with differing standards, policies and procedures.

## C.3 Service and Subject-specific Portals

Service-based portals are now very common. Examples include Google, Amazon and e-Bay which are familiar to millions of people worldwide. They have many similarities to project portals, but are focussed on the end to end delivery of a specific service or set of services to its customers/ users.

There are many subject-specific portals, such as Arxiv `http://arxiv.org` (Cornell University), PubMed `http://www.pubmed.com` (NIH), or UKPMC: UK PubMed Central `http://www.wellcome.ac.uk/doc_WTD015366.html` (Wellcome Trust). Many experienced researchers prefer subject-specific portals which contain deep-search and other facilities which they can use based on specilist vocabulary and subject knowledge.