

JISC Defining Image Access Project

Final Report

Images and Repositories: Present Status and Future Possibilities

EXECUTIVE SUMMARY

Authors: **David Shotton** david.shotton@zoo.ox.ac.uk
 Jun Zhao jun.zhao@zoo.ox.ac.uk
 Graham Klyne graham.klyne@zoo.ox.ac.uk

The Image Bioinformatics Research Group
Department of Zoology, University of Oxford
South Parks Road, Oxford OX1 3PS, UK

This version of Executive Summary: Final 16 August 2007

Citation to use when referring to the Final Report from which this Executive Summary is extracted:
Shotton, D.M., Zhao, J and Klyne, G. (2007). Images and Repositories: Present Status and Future Possibilities. Final Report of the JISC *Defining Image Access* Project (January – June 2007).

This report is downloadable from the JISC *Defining Image Access* Project web page
(http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres/defining_image_access.aspx)
and from the Oxford Research Archive (<http://ora.ouls.ox.ac.uk>).

Copyright © 2007 David Shotton, Jun Zhao and Graham Klyne.

Published under the Creative Commons Attribution-Noncommercial-Share Alike 3.0 License
(<http://creativecommons.org/about/licenses/meet-the-licenses>).

EXECUTIVE SUMMARY

Abstract

The **JISC *Defining Image Access* Project** was a six-month requirements analysis project (January to June 2007) funded by the JISC to investigate the feasibility of creating data webs that would permit subject-specific search integration of institutional repository image collections using Semantic Web techniques. This **Final Report**:

- Describes the concept of data webs, in contrast to other forms of data integration across distributed heterogeneous resources;
- Describes project evaluations (a) of the institutional repositories at Cambridge, Imperial College, Oxford and Southampton Universities in terms of their software, image holdings and metadata exposure mechanisms, (b) of related projects, and (c) of Web standards, tools and software applications that might be employed to construct a data web for research images;
- Reports eight conclusions from these investigations, and proposes the future development of a demonstrator image web based on these findings and our pilot software developments; and
- Makes ten recommendations to institutional repository managers and to the JISC.

Summary of project achievements

As a potential solution to the problem of locating data scattered across heterogeneous resources, we have proposed the development of subject-specific data webs (<http://www.rin.ac.uk/data-webs>) that use the Web as their native platform and enable integrated access to images or other datasets relating to these particular subjects. Within each data web, loosely coupled software services will be used to combine metadata describing research datasets in distributed resources, in a manner that permits discovery and provide links back to the original data sources to allow data delivery.

We started our work on data webs from the premise that much useful research data is presently unpublished and could usefully be published on the Web, and that lightweight Web-based tools could be used to link these diverse publications into more or less coherent bodies of research information for various domains of interest. The mandate of the *Defining Image Access* Project was specifically to examine research images in institutional repositories across all subject domains, and to explore the feasibility of creating data webs to link subject-specific images from different repositories.

During this project, we established a significant core body of knowledge and expertise concerning Web-based standards, tools and technologies available to create data webs, and about the images in institutional repositories and the problems and opportunities associated with integrating them. Our findings were recorded as the project progressed in a project wiki Web site ([http://imageweb.zoo.ox.ac.uk/wiki/index.php/Defining Image Access](http://imageweb.zoo.ox.ac.uk/wiki/index.php/Defining_Image_Access)). This has become a valuable resource, ranking surprisingly high in relevant Google searches, vindicating our wiki philosophy to make all our project findings immediately and publicly accessible. We saw growth of interest in our data web concept from people outside the project, which has led to its inclusion in two non-JISC project proposals, one in the arts and the other in the sciences.

We also conducted four small but very productive project workshops, and held individual meetings with repository partners, which permitted us to learn from other projects and publicize our activities. Throughout the project, we met and exchanged ideas with key people involved in related JISC projects, and learned how we might integrate our activities with theirs.

Our findings reinforced our view that the Web-as-platform approach to data integration is feasible and widely applicable, and permitted us significantly to refine our ideas about data web functionality. The availability of several mature tools supports our present idea of using SPARQL as a central technology for accessing diverse information sources.

By creating an Eprints repository for publication of research images and metadata from *Drosophila* gene expression research undertaken by colleagues here in Oxford, we showed that existing repository software can be adapted for wider use.

We constructed a plan to create an exemplar data web to provide interoperability between domain-specific repository journal articles and relevant research datasets located elsewhere. This plan takes account of the lessons we learned through the conduct of this project, and as such allowed us greater opportunity to evaluate and mitigate risks that would have been inherent in an earlier attempt to create a data web solely over repository holdings.

Principal conclusions

- C1:** Institutional repositories should be seen as just one element in a wider ecosystem of Web-based publication of research data and scholarly writings, that also includes research group databases, national repositories and global databases. The Web, and Web-standard technologies, must be recognised as the primary mechanisms for bringing together these different sources. Our vision of a data web is an element of this view, using Semantic Web standards and tools to combine information from disparate sources for access by both human readers and computer software.
- C2:** Institutional repositories currently contain few image collections, these image collections mostly lack adequate domain-specific metadata, and existing repository interfaces are not well equipped to serve domain-specific metadata in a machine-readable manner. These limitations indicate a need for some preparatory work, particularly in terms of image submissions to, and access from, institutional repositories, before our initial idea for the creation of subject-specific inter-repository image webs becomes an achievable goal.
- C3:** Through our work with EPrints, we have shown that it is possible to adapt repository software for research group data and metadata publication. We anticipate that by using existing repository software in this manner, eventual data migration to institutional repositories will be facilitated, extending the benefits of such repositories to research groups.
- C4:** If repositories are to be widely used to house research data, attention also needs to be directed to tools that support the gathering of appropriate metadata as an early activity within the research process, in advance of the time of its eventual publication. Such tools should augment current research practices rather than becoming an imposition upon them.
- C5:** Given the current state of institutional repository holdings, we believe that data webs would at present be more usefully deployed to link the journal articles and papers that presently constitute the bulk of such holdings with the research datasets and images upon which these articles are based, located elsewhere. Such data webs would complement service frameworks that facilitate metadata capture at the time of research image creation, and that enable Web publication of the images and their metadata.
- C6:** A number of mature software tools are available, based on Semantic Web technologies, that provide key elements of functionality needed to implement a data web. Such data webs should comprise independent loosely coupled light-weight services for schema registration, co-reference resolution and distributed query processing.
- C7:** Nevertheless, building a data web remains a significant implementation task, and the sources across which such data webs operate need to be carefully chosen. The proposed schema registry and co-reference services for each data web serving a particular knowledge domain will require hand-crafted alignment. Thereafter, handling of instance metadata will be automatic.
- C8:** Additional Semantic Web tools created by other research groups, such as mSpace or jSpace, seem well suited to provide semantic discovery services over data webs. Furthermore, a semantic tagging service such as RichTags presents a promising mechanism to facilitate user annotation for *post hoc* addition of metadata, for example, relating the original research to other areas of interest.

Recommendations for repository managers and the JISC

- R1:** The value of institutional repositories could be enhanced by facilitating the deposition of, and subsequent programmatic access to, image collections and other datasets with appropriate domain-specific metadata. The availability of suitable tools and standards to support this is

currently patchy, and the appropriate mechanism may vary depending on the repository software used. We recommend that repository managers seek out suitable tools and explore the adaptation of existing tools to support repository data deposition and access, and articulate to tool developers the requirements and constraints under which such tools must operate, including mechanisms for bulk ingest of data collections and storage of arbitrary domain-specific metadata, and the provision of search and browse interfaces that take account of the nature of the data type (e.g. images, media clips) when presenting query results.

- R2:** We recommend that repository managers should develop facilities and policies to build researchers' confidence that institutional repositories can keep their data safe and maintain the confidentiality of information relating to work in progress.
- R3:** The paucity of available image metadata suggests that mechanisms for *post hoc* annotation of published images and datasets, with appropriate provenance records, will be required, if such images are to be useful for re-use in new lines of research. It is not clear to what extent such facilities should be provided by institutional repositories but we recommend that, at the least, repository managers should allocate stable URIs for published images and image collections (possibly URNs or DOIs), so that reliable third party annotation systems can be deployed.
- R4:** As and when suitable tools are available, we recommend that repository managers should deploy SPARQL endpoints to supplement OAI-PMH as a means for machine-mediated discovery of repository holdings based on metadata queries, and should facilitate exposure of additional metadata, beyond the usual Dublin Core elements.
- R5:** We applaud the work to create lightweight, common submission mechanisms and repository interoperability protocols (e.g. ORE and SWORD), and we recommend that the JISC works with repository administrators, users and tool developers to ensure that a single act of submission is all that is required to deposit data and supporting metadata to multiple sites.
- R6:** We recommend that the JISC updates its Repositories Roadmap and Information Environment Architecture documents to present the IE as an overlay on the Web-as-platform, with recommendations for lightweight service-oriented architectures that employ the Web as the platform, that encourage the use of Semantic Web and Web 2.0 technologies where appropriate, and that aid integration with external non-JISC IE components (Powell, 2007).
- R7:** We recommend that the JISC ensures that the Application Profile for Images is not limited to (a) Dublin Core–FRBR type metadata, but also includes (b) regulatory metadata defining IPR, copyright and conditions for reuse, (c) structural metadata relating to file size, format and encoding, (d) versioning and provenance metadata, and (e) semantic metadata describing the content, meaning and significance of the images (Shotton *et al.*, 2002).
- R8:** We recommend that the JISC encourages innovative interactions between JISC projects related to research data and images, and strives to engage researchers and research tool developers, in addition to members of the repository development and library communities.
- R9:** We recommend that the JISC commissions and funds a competent group of experts such as the JISC Common Repository Interface Working Group to create SPARQL endpoints for all commonly used institutional repository software systems.
- R10:** Finally and most importantly, as we move rapidly into an era of data-driven research and scholarship (Lyon, 2007) (<http://www.rin.ac.uk/data-publication>), in which effective data management will be essential to maintain research competitiveness, we recommend that the the JISC should be proactive in funding research projects that (a) assist researchers in capturing descriptive information about research datasets (i.e. domain-specific metadata) as early as possible in their workflows in ways that enhance existing research practices, (b) facilitate the submission of such semantically enhanced research datasets to open access repositories; and (c) promote the accessibility to and reuse of research data, by the creation of data webs or similar services that provide interoperability between institutional repositories and third-party data resources, and that enhance the links between research publications and the primary datasets upon which they are based.

References

- Lyon (2007). Dealing with data: roles, rights, responsibilities and relationships. JISC commissioned report,
http://www.jisc.ac.uk/media/documents/programmes/digital_repositories/dealing_with_data_report-final.pdf.
- Powell (2007). The JISC Repository Roadmap - are we heading in the right direction? JISC Repositories Conference, Manchester, <http://www.slideshare.net/eduservfoundation/the-repository-roadmap-are-we-heading-in-the-right-direction>.
- Shotton, Rodriguez, Guil *et al.* (2002). A metadata classification schema for semantic content analysis of videos. *Journal of Microscopy* **205**: 33 -42.