

JISC Defining Image Access Project

Final Report

Images and Repositories: Present Status and Future Possibilities

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

Phone: +44 (0) 1865 271193.

With two appendices:

Appendix A: An independent commentary on the *Defining Image Access* Project.

by **Julie Allinson** j.allinson@ukoln.ac.uk

Repositories Research Officer, UKOLN, University of Bath, Bath BA2 7AY, UK

Appendix B: A Scholarly Works Application Profile, conforming to the standard established by (Allinson *et al.*, 2007) by **Julie Allinson** and **David Shotton**

This version of the Final Report

FINAL

17 August 2007

Citation to use when referring to this report:

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).

Final Report and its Executive Summary are available as PDF files, or may be viewed as HTML in a Web browser. The Scholarly Works Application Profile, providing metadata describing this Final Report, is in machine-readable XML format.

The PDF and XML files are accessible 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>). These and the HTML version are available at <http://imageweb.zoo.ox.ac.uk/pub/2007/DefiningImageAccess/FinalReport/>.

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

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

1. TABLE OF CONTENTS

1.	TABLE OF CONTENTS	2
2.	EXECUTIVE SUMMARY	5
2.1.	Abstract	5
2.2.	Summary of project achievements	5
2.3.	Principal conclusions	6
2.4.	Recommendations for repository managers and the JISC	6
3.	ADMINISTRATIVE DETAILS	8
3.1.	JISC Project title	8
3.2.	Project purpose	8
3.3.	Project team	8
3.4.	Project consultant partners	8
3.5.	Project duration, funding and management	8
4.	PREFATORY REMARKS	9
5.	ACKNOWLEDGEMENTS	9
6.	BACKGROUND	10
6.1.	The importance of images	10
6.2.	Previous approaches to the integration of heterogeneous data resources	11
6.3.	The data web philosophy	12
7.	OVERVIEW OF THE <i>DEFINING IMAGE ACCESS</i> PROJECT	15
7.1.	Motivation	15
7.2.	Abstract from JISC <i>Defining Image Access</i> Project grant application	16
7.3.	Project description	16
7.4.	Aims and objectives	16
7.5.	Approaches	17
7.6.	Project deliverables	18
8.	PROJECT ACTIVITIES	19
8.1.	Project meetings	19
8.2.	Survey work	19
8.3.	Meetings with repository partners	19
8.4.	Software evaluation	20
8.5.	Technical design activities	20
8.6.	Dissemination activities	20
9.	REPOSITORY SURVEY FINDINGS	22
9.1.	Evaluation of repository systems	22
9.1.1.	DSpace	22
9.1.2.	Fedora	22

JISC <i>Defining Image Access Project</i> Final Report	Shotton, Zhao and Klyne, 2007	3
9.1.3. EPrints		22
9.1.4. Conclusions		23
9.2. Evaluation of repository image holdings		23
9.2.1. Dspace@Cambridge		23
9.2.2. The Southampton EPrints repository		24
9.2.3. Imperial College		24
9.2.4. University of Oxford		24
9.2.5. Conclusions		24
9.3. Evaluation of standards and protocols for repository metadata exposure		25
9.3.1. Metadata standards		25
9.3.2. Dublin Core		26
9.3.3. CIDOC CRM Core		26
9.3.4. Domain-specific metadata schemas and ontologies		26
9.3.5. OAI-PMH		27
10. SOFTWARE TOOLS EVALUATION		27
10.1. Commentary on Web and Semantic Web standards		27
10.1.1. Web standards		27
10.1.2. Semantic Web standards		28
10.1.3. Comparison of SPARQL and OAI-PMH		28
10.2. Tools for building data webs		29
10.2.1. Tools for annotation		29
10.2.2. Tools for faceted semantic browsing		29
10.2.3. Tools for creating SPARQL endpoints over OAI-PMH		30
10.2.4. Tools for creating core data web functionality		30
11. SURVEY OF RELATED R&D PROJECTS		31
11.1. CAIRO		31
11.2. CLADDIER		31
11.3. Common Repository Interface Working Group (CRIG)		32
11.4. DExT		32
11.5. Dictate		32
11.6. eBank-UK, R4L, SPECTRa		32
11.7. ImageStore project		33
11.8. Information Environment Metadata Schema Registry (IEMSR)		33
11.9. Intute Repository Search		33
11.10. OAI-ORE		33
11.11. Rich Tags		34
11.12. StORe		34
11.13. SWORD		34
12. PROJECT SOFTWARE DEVELOPMENTS		35
12.1. Creation of an EPrints repository for capturing research images and metadata		35
13. AREAS REQUIRING FURTHER EVALUATION		37
13.1. Metadata acquisition		37
13.2. User interfaces		37
13.3. Ontologies and annotation		37
13.4. Schema alignment and co-referencing		37

JISC <i>Defining Image Access Project</i> Final Report	Shotton, Zhao and Klyne, 2007	4
13.5. Metadata pre-harvesting versus distributed querying		38
14. CRITIQUE OF THE PROJECT		38
14.1. Summary of project achievements		38
14.2. Project shortcomings		38
15. PROPOSAL FOR FUTURE WORK		39
15.1. A design outline for an image data web as part of an integrated workflow for research image data publication and reuse		39
15.2. Software approaches to achieving a service-oriented architecture		40
15.3. Data web software framework		40
15.4. Implementation plan		41
15.5. Creation of a generic SPARQL endpoint for OAI-PMH repositories		42
16. PROJECT CONCLUSIONS AND RECOMMENDATIONS		42
17. GLOSSARY OF NAMES, ABBREVIATIONS AND ACRONYMS		43
REFERENCES		46
18. INTRODUCTION TO APPENDIX A		48
19. APPENDIX A: AN INDEPENDENT COMMENTARY BY JULIE ALLINSON		49
1. Images in an institutional repositories context		50
2. The data webs concept		51
3. Joining up with JISC services, projects and activities		51
4. Discovery services		51
5. Linking research data and publications		52
6. Repositories-related projects		53
Existing projects:		53
New projects:		53
7. JISC Services		55
Sources of content		55
Machine services		55
Advisory services		55
8. Beyond JISC, connections with the wider world		56
9. Data webs and interoperability		56
The JISC Information Environment		56
The JISC/DEST E-Framework for Research and Education		57
10. Metadata and object modelling issues		57
11. Conclusion		58
20. APPENDIX B: THE SCHOLARLY WORKS APPLICATION PROFILE FOR THE DEFINING IMAGE ACCESS PROJECT FINAL REPORT		60

2. EXECUTIVE SUMMARY

2.1. 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.

2.2. 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.

2.3. 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.

2.4. 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.

3. ADMINISTRATIVE DETAILS

3.1. JISC Project title

Defining Image Access: Requirements for interoperable discovery and delivery of image data stored in DSpace, EPrints and Fedora-based institutional repositories using a data web approach

3.2. Project purpose

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

3.3. Project team

David Shotton Principal investigator (david.shotton@zoo.ox.ac.uk),
Graham Klyne Project manager (graham.klyne@zoo.ox.ac.uk),
Jun Zhao Postdoctoral computing officer (jun.zhao@zoo.ox.ac.uk).

3.4. Project consultant partners

Julie Allinson UKOLN, Digital Repositories Programme Support Team (<http://www.ukoln.ac.uk/repositories/digirep/>).
Dan Brickley Independent Semantic Web consultant (<http://danbri.org/>).
Michael Fraser and Peter Robinson Learning Technologies Group, Oxford University Computing Services (<http://www.oucs.ox.ac.uk/>).
Dolores Iorizzo and Yiota Polydoratou Imperial College Internet Centre and Imperial College Library (<http://www.internetcentre.imperial.ac.uk/>).
Jessie Hey University of Southampton School of Electronics and Computer Science (<http://www.ecs.soton.ac.uk/>) and e-Prints software team (<http://www.eprints.org/software/>).
Neil Jefferies and Sally Rumsey Oxford Library Services (<http://www.ouls.ox.ac.uk/>).
Patricia Killiard DSpace@Cambridge, Cambridge University Library (<http://www.dspace.cam.ac.uk/>).
Brian Matthews STFC (formerly CCLRC) e-Science Centre, Rutherford Laboratory (<http://www.e-science.clrc.ac.uk/>).
David Wallom Oxford e-Research Centre (<http://www.oerc.ox.ac.uk/>).

3.5. Project duration, funding and management

Period: January 2007 to June, 2007 inclusive, extended to July 2007.
Funding: From the JISC: £62,991. Funded from the Discovery to Delivery strand of the JISC Repositories and Preservation Programme.
Programme manager: Balviar Notay (b.notay@jisc.ac.uk).

4. PREFATORY REMARKS

This report differs somewhat from conventional print-based JISC project reports, in that it contains many active hyperlinks to sections within the JISC *Defining Image Access* Project wiki at [http://imageweb.zoo.ox.ac.uk/wiki/index.php/Defining Image Access](http://imageweb.zoo.ox.ac.uk/wiki/index.php/Defining_Image_Access), where further detail is provided (These hyperlinks can be converted into full URLs by the addition of the prefix “<http://imageweb.zoo.ox.ac.uk/wiki/index.php/>”). Use of these hyperlinks has enabled us to keep the length of the report itself within manageable bounds, while still providing access to more detailed information on individual topics that might be of particular interest to certain readers.

Additionally, since this report has been largely derived from the content of our project wiki, which was used throughout the project as the repository within which all project-related information was incrementally accumulated, we have retained the personal style used there, which represents the collective view of the members of the Image Bioinformatics Research Group (IBRG) at the University of Oxford.

With the exception of the few formally citable publications (e.g. Van de Sompel *et al.*, 2005) that are given in full in the concluding list of references, citations of and references to others’ work and projects take the form of URL hyperlinks to external web sites, or to relevant sections of our own project wiki, embedded within the text of this report.

The Project wiki, from which this information will continue to be made freely available, has become an information resource that now ranks surprisingly high in relevant Google searches. It is our intention to continue to support and develop it for the foreseeable future beyond the termination of the JISC *Defining Image Access* Project, its content being extended, enriched and updates as appropriate.

However, since departmental and project-specific Web servers have no assurance of permanence beyond the next decade, we are currently negotiating with the JISC to determine how and where best to archive the current (summer 2007) version of the Defining Image Access pages from our ImageWeb wiki, so that the data contained within it will remain accessible in the long term. Notification of the location of this archival version will be placed on the JISC Defining Image Access Project web page (http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres/defining_image_access.aspx), available for use if the live wiki ceases to be accessible.

[Note: In this report, as in common usage among members of the computing community, the word ‘metadata’ is used as a collective noun taking the singular form of verbs. Furthermore, the plural form used for the word ‘schema’ is ‘schemas’ rather than ‘schemata’. We apologize in advance to Latin purists to whom this may cause offence.]

5. ACKNOWLEDGEMENTS

The authors acknowledge with gratitude the significant contributions of time and ideas by the project’s Consultant Partners listed in Section 3.4, and by the external experts invited to the project’s four workshops, whose names and contributions are detailed in the meeting reports at [http://imageweb.zoo.ox.ac.uk/wiki/index.php/Defining Image Access#Project meetings](http://imageweb.zoo.ox.ac.uk/wiki/index.php/Defining_Image_Access#Project_meetings). We particularly wish to acknowledge that our data web ideas have been improved by ideas flowing from private discussions with Brian Fuchs of the Imperial College Internet Centre, and Martin Doerr of the Institute of Computer Science, Foundation for Research and Technology – Hellas (ICS FORTH; <http://www.ics.forth.gr/>), the principal developer of CIDOC CRM, to both of whom we are most grateful. We are most grateful to Julie Allinson and Dan Brickley for their comments on earlier drafts of this Final Report, and to Julie Allinson for her insightful commentary on the project (Appendix A) and for the Scholarly Works Application Profile (Appendix B). Finally, we wish to express our gratitude to the JISC for funding this investigation, and particularly to Balviar Notay, our JISC Programme Officer, for her support and guidance throughout the project.

6. BACKGROUND

6.1. The importance of images

Research results may be usefully, if crudely, divided into two kinds: ‘universals’ and ‘particulars’. Certain limited types of research data, such as the sequence of a gene or the 3D structure of a protein are, to a first approximation, universal truths. They need only be discovered once, form bounded data sets, and are seen as fundamental knowledge to which all should have access. By community consent, and with significant central funding, they are published in global bioinformatics databases mirrored in Europe (at the European Bioinformatics Institute, Cambridge), in the United States and in Japan. In contrast, ‘particulars’ are specific observations that vary between experiments and research groups, for example 3D images of archaeological discoveries, simulations of airflow over an aircraft wing, videos of social science interviews, or assay results for gene expression. (In the broader, non-research context, particulars are also exemplified by recordings of particular items or events, such as press photographs, concert recordings, micrographs of cells, and wildlife films). While constituting important parts of the research record in the arts, sciences and humanities, particulars form unbounded datasets that can never be regarded as complete. Publications of all such particulars to global databases is both inappropriate and impractical. Unfortunately, the majority of such research data at present go unpublished. Instead, we believe that publication of such datasets, and particularly of images of all types, should be made in local open access institutional repositories or research databases, and should be accompanied by sufficient domain-specific metadata to permit their discovery by digital searches. It should be clearly recognised that datasets submitted to databases or repositories with insufficient metadata are being consigned to costly digital data graveyards from which resurrection is likely to be extremely difficult.

Images, in particular, play a vital role in academic research and teaching, and their acquisition is often costly and time consuming. Furthermore, the storage requirement for such images is large, particularly if they are multidimensional. However, despite popular misconception, the central problem with images is not their size or dimensional complexity, but that, unlike text documents and certain forms of scientific data (e.g. gene sequences), they are generally not self-describing. While they may be easily interpreted by humans, they typically have no internal semantics that are readily extractable by present computing technologies, notwithstanding recent advances in machine vision research. Descriptive metadata about the images is therefore essential to bridge this ‘semantic gap’.

Even with good metadata, finding images is not easy, since different data sources exhibit varying degrees of syntactic and semantic incompatibility, making it impossible to find relevant images without searching each source individually. Nevertheless, the benefits of doing so could be considerable, including the ability to study more of the images upon which a particular research publication is based than are included in the article’s figures, and the potential of undertaking meta-research made possible by easy access to images from various sources.

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>; <http://imageweb.zoo.ox.ac.uk/wiki/index.php>) that use the Web as their native platform and enable integrated personalized user access to images or other dataset relating to these particular subjects. Within each data web, integration of the heterogeneous distributed resources is achieved by semantic mappings of their individual metadata schemas to a core schema. Lightweight software tools are then used to access metadata describing research images in distributed sources, and combine this into a single searchable RDF graph that permit discovery and provide links back to the original data sources to allow data delivery.

6.2. Previous approaches to the integration of heterogeneous data resources

The term “data web” has previously been used, for example for systems that integrates data from the major bioinformatics databases (Grossman, 2003). However, the technologies hitherto proposed to provide interoperability between distributed databases have been centred on more heavyweight Web and Grid Services, have used a top-down resource federation model, and have not employed Semantic Web technologies.

One example of this within the UK e-Science framework is the OGSA-DAI development (Open Grid Services Architecture – Database Access and Integration; <http://www.ogsadai.org.uk>) for distributing queries across participating databases. This has involved large teams of developers, has consumed huge amounts of the e-Science budget, and has resulted in a heavyweight software system that, while prototyped in grid environments, is at present little used by the average bench scientist.

Another similar UK e-Science project was the eDiaMoND Project (<http://www.ediamond.ox.ac.uk/project.html>) that set out to integrate mammography data into a single national database. Although most of the problems in achieving that goal were found to be social rather than technical, having to do with confidentiality of patient records and variations of interpretive practice between different hospitals, it is interesting to note that the follow-on GIMI Project (<http://www.gimi.ox.ac.uk/applications.html>), funded by the DTI Technology Programme, made the decision to drop the use of such heavyweight Grid Services.

In the United States, the CORDRA Project (<http://www.cordra.net/>) has specified an aggregation service similar in concept to our data web. The related aDORe Project (Van de Sompel *et al.*, 2005) has demonstrated many of the CORDRA concepts in practice, and from these has developed the current OAI-ORE Project discussed in 10.6.10 below.

We believe that the lightweight data web approach, involving an initial search across core metadata followed by linking to the source data, will prove to be a faster and more scalable method of data access across multiple content repositories than those mentioned above. While others have elegantly demonstrated the possibility of using Semantic Web technologies to integrate information from bioinformatics databases direct to the user’s desktop (Neumann and Quan, 2006), no-one has previously suggested their use to create independent data webs that integrate third party information into specialist publicly searchable metadata registries. Furthermore, this solution to the integration needs of an ever increasing volume of research data is not envisaged in either of the recent reviews of the future of computing (Emmott, 2006), (Mugleton *et al.*, 2006).

Nearer in concept to our data web approach is the ‘data space’ proposals of Halevy and his colleagues (Franklin *et al.*, 2005; Halevy *et al.*, 2006), although these are made very much from the perspective of the traditional relational database community, do not involve Semantic Web technologies such as RDF for data integration, and were proposed before the development of SPARQL as the universal RDF query language.

Even closer in concept, although again not using Semantic Web technologies, is the service specific for images called PictureAustralia (<http://www.pictureaustralia.org/>), a service hosted by the National Library of Australia. This service provides a single access point to the digitised pictorial collections of a range of participating individuals or agencies, providing images of people, events and places in Australia, such as photographs of the Sydney Harbour Bridge, and of Australian artwork and objects. The PictureAustralia service harvests Dublin Core metadata from the agencies on a monthly basis, using OAI-PMH (see Section 9.3.5 below) or HTML screen-scraping for the smaller agencies, and creates a single XML keyword index that users can search. Search returns are illustrated by thumbnail images that are not stored centrally but rather are dynamically downloaded to the user’s browser from the original data sources. From such search returns, users can use

hyperlinks to gain access to the original full-size images from the publishing sites, for a fee if necessary. Copyright and access rights to all images and thumbnails thus continue to be held by the image providers. PictureAustralia is represented in the photo-sharing Web site Flickr <http://www.flickr.com/people/92276616%40N00/>. The ease of operation of PictureAustralia has much to teach us when we come to implement an image web, and its free-to-users advertisement-free business model is also of interest. Further details are given at <http://imageweb.zoo.ox.ac.uk/wiki/index.php/DefiningImageAccess/Project/PictureAustralia>.

One of the key elements in our Semantic Web approach to data integration over heterogeneous repositories is that of schema alignment, which ensures that alternative metadata descriptors used by different data providers (e.g. “author”, “creator”, “photographer”) are recognised as referring to the same concept. Good work in this area has already been undertaken by Martin Doerr and his colleagues (Kondylakis *et al.*, 2006) in conjunction with his development of the CIDOC Conceptual Reference Model (see Section 9.3.3 below), and this has proved helpful to us (see <http://imageweb.zoo.ox.ac.uk/wiki/index.php/DefiningImageAccess/Resource/SchemaAlignment>).

6.3. The data web philosophy

Tim Berners-Lee, the inventor of the World Wide Web, has argued from its inception for a distributed “web of data” in which all Web-accessible information carries its own machine-readable semantics (Berners-Lee, 1989; Berners-Lee, 1999; Berners-Lee, 2007). We propose a focused implementation that offers a step towards Berners-Lee’s vision of the Semantic Web (<http://www.w3.org/2001/sw/>), namely the creation of subject-specific data webs to meet defined research data publication, access, integration and meta-research needs. The data web concept rests on the fundamental observation that distributed metadata related to a particular subject can be integrated, both syntactically and semantically, if it can first be mapped to a common core data model and represented as RDF (<http://www.w3.org/RDF/>), since, as eloquently stated by (Connolly, 2007), with RDF “URI-based data merging is built in”.

We believe that, with a few specific exceptions, the standards and technologies that will allow us to create data webs, such as simple HTTP web protocols in the Representational State Transfer (REST) style (Fielding, 2000), already exist and have been proven in operation at Internet scale. The basic components of a data web are shown in Figure 1.

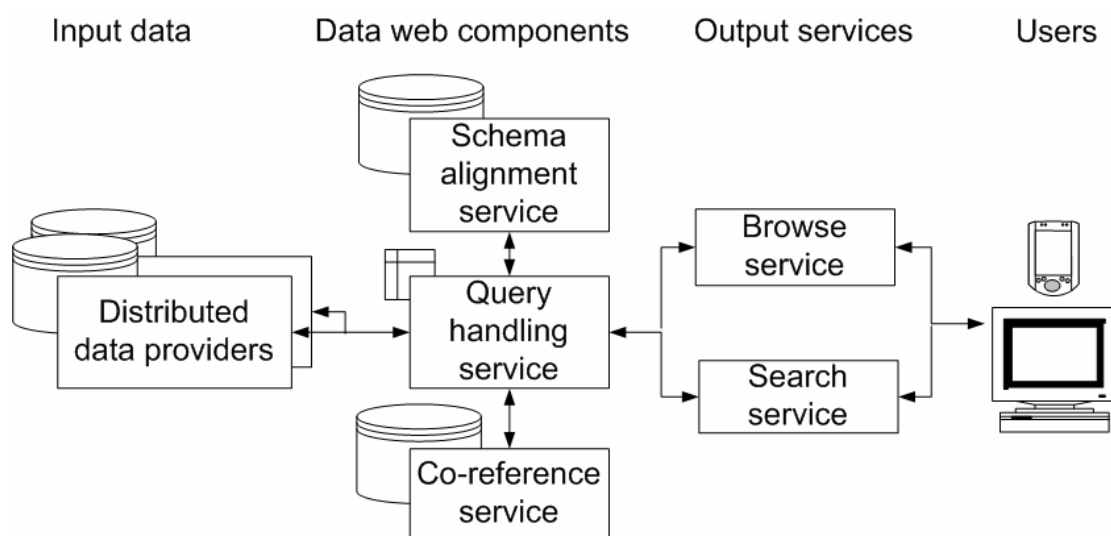


Figure 1. The fundamental components of a data web

Data webs do not require all the data provided by third-party resources to be semantically coordinated, or constrained to conform to a single externally imposed model of information management, but rather permits them to maintain their unique characters and continue independent

publication of information describing their holdings. Using the premise that a little semantics goes a long way, interoperability is instead enabled by creating a semantically integrated view over the heterogeneous data within the distributed data repositories, by bespoke mapping of their independent database schemas to a common representation. *Separately for each data web serving a particular knowledge domain*, this data integration occurs in three stages:

- First, a core schema specific to the domain of interest is created, defining the basic vocabulary that defines the domain and will be used to search the data web.
- Second, in separate acts of ‘subscription’ to the data web, the independent metadata schema of each subscribing data source is mapped to the data web’s core schema, this mapping being recorded within the Schema Alignment Service.
- Third, any known co-references to data objects identified by elements of the data web are recorded within the data web’s Co-reference Service.

These initial activities require hand crafting by either data source or data web personnel. Subsequently, key metadata elements describing particular digital objects (‘instances’) within the data sources can be accessed via the data web and integrated in an entirely automated process.

Depending upon the technical nature of the data source, a software ‘adapter’ may need to be installed, enabling metadata elements describing particular digital objects (instances) within the data source to be obtained by the data web’s central Query Handling Service. Ideally, this adaptor should provide a **SPARQL endpoint** (http://wiki.ontoworld.org/index.php/SPARQL_endpoint) so that a SPARQL query from the data web can retrieve an RDF description of data source holdings, enabling resource discovery, as described in Sections 10.1.2 and 10.2.3 below. We have identified two possible designs for such adapters: (1) local metadata harvesting and rewriting to an intermediate RDF metadata store, against which SPARQL queries are resolved, or (2) query rewriting from SPARQL into the native query format of the resource (e.g. SQL) (Figure 2).

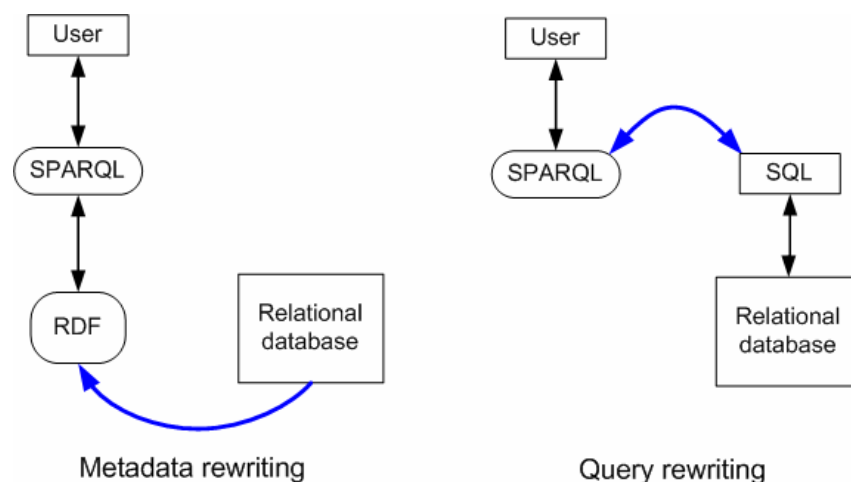


Figure 2. Alternative types of SPARQL endpoint

By presenting data as RDF through SPARQL, the data web overcomes syntactic differences between data providers, while by applying information from the Schema Alignment and Co-reference Services, the data web Query Handling Service can resolve semantic differences between the various data sources.

The browse and search components of the data web (Figure 1) present Web browser interfaces, and transmit user queries to the data web’s Query Handling Service, which retrieves RDF metadata describing the query subject from across the data web. The aggregated responses can then be presented to the user, who is able to select one or more of the returned values and link directly to the data sources to obtain access to the original data. The data web thus acts as a data marshal, ordering and integrating the data web participants’ holdings into a single searchable RDF graph,

providing the basis for both human and programmatic query access to the data from all sources covered by the data web.

In principal, different data webs might access the same data sources for quite distinct purposes, for example accessing cellular images either to compare microscopy techniques or to study disease progression. Each data web will provide a focused service to meet defined data access, integration and meta-research needs of a specific community of interest. A particular data web will thus embody domain-specific knowledge, even though its software components may be quite generic.

The architecture of a data web is such that additional data providers can be subscribed to the data web at any time, thereby enriching the total content, by presenting to the Schema Registry a mapping of their schemas to the core schema, and (if necessary) by presenting to the Co-reference Service a mapping of any alternative identifiers used to reference common entities.

It should be noted that in this model only descriptive metadata is harvested and made available to the user by the data web. Access to and copyright of the data objects themselves is left firmly in the hands of the original data providers.

Data webs of this type will have all the advantages of the World Wide Web itself, namely distributed data, freedom and decentralization of publication, a “missing isn’t broken” open world philosophy (Brickley, 2003), lack of centralized control, evolvability, and scalability. However, unlike the Web as a whole, each particular data web will provide tailored access to just one bespoke information market, with the following advantages over Web search engines such as Google:

- It permits access to the ‘Deep Web’ of database content that search engines conventionally cannot access (He *et al.*, 2007).
- By being specifically targeted to a particular knowledge domain or context, it achieves a significantly higher signal-to-noise ratio within search returns.
- It involves integration of information with ontological underpinning, semantic coherence, and truth propagation, so that the truth of the integrated information is logically entailed by the truth of the components (see <http://en.wikipedia.org/wiki/Entailment>).
- Perhaps most importantly, it emphasises programmatic access, enabling added-value services to be build on top of one or more data webs.

This provision of third-party programmatic access enables the information aggregated by a data web to be reused and subjected to more detailed semantic processing in unanticipated ways by future services and mash-ups, thereby adding value. Within the biological realm, examples of such output-level added-value services for images could include semantic slicing through large datasets to present users with topic-specific views (e.g. all the images relating to a specific gene across different species), and mashups of geospatial metadata with Google Maps for localization of wildlife photos.

This model of a semantically-enabled service-oriented data web infrastructure is generic, being applicable to other data types than images. It supports Open Source access, and integrates well with a local Web publication paradigm.

Data webs built across image repositories will benefit the primary data holders by making their holdings cross-searchable and by bringing additional users to their sites, without in any way controlling or constraining access to the primary data. These remain under the full copyright and access control of their source repositories, while their metadata remain available on the Web for use by other presently unforeseen applications including novel data mining and analysis services.

In this way, we hope that the holdings of the institutional repositories with whom we are currently working will become more integral components of the day-to-day information environment of academic researchers, teachers and students, and even the general public.

7. OVERVIEW OF THE *DEFINING IMAGE ACCESS PROJECT*

7.1. Motivation

The key motivation of data webs is the following: A lot of data is being published on the Web in an uncoordinated, even chaotic, fashion. Data sharing communities are being formed, but they tend to be closed in the sense of depending on conformance to some private community schema. The data web philosophy is to tap into and extend the reach of these web data communities without requiring them to change their existing practices.

Realization of interoperable access to the world's scholarly information requires semantic integration of the distributed, heterogeneous and presently non-interoperable digital resources. Within the Image Bioinformatics Research Group at the University of Oxford, our ImageWeb work has been motivated primarily by the need to provide access to and facilitate reuse of research images created as part of ongoing post-genomic life science research.

Making the image web vision a reality has involved a number of identifiable phases. Phase One (late 2005) involved developing our initial concepts. Phase Two (2006) saw the initial dissemination and review of the concepts, particularly through the First Research Information Network Workshop (Imperial College, June 28, 2006) entitled *Data Webs: new visions for integrating research data on the Web*, formation of the BioImageWeb Consortium, and fundraising activities. Phase Three (Jan – June 2007) has involved requirements analyses, made possible by the JISC-funded *Defining Image Access Project*, for which this is the Final Report. Phase Four, which will involve creation of a demonstrator data web for a specific area of research activity, will (funding permitting) commence in October 2007, leading to more general deployment of data webs in future years.

The BioImageWeb Consortium is an informal group of parties interested in taking forward the image web vision and making it a reality. Led by the Image Bioinformatics Research Group at the University of Oxford, this open consortium currently comprises representatives from:

- Leading publishers of subscription and open access research journals (Elsevier, Nature, Oxford Journals and Wiley-Blackwell; BioMed Central and the Public Library of Science);
- The institutional repository Consultant Partners of this project, namely those of the universities of Cambridge, Imperial College, Oxford and Southampton;
- Other stakeholders, including CCLRC (now STFC) and UKOLN (both also Consultant Partners of this project), and the British Library, ILRT, CrossRef, SPARC Europe and Ingenta (now Publishing Technology plc); and
- Professional researchers and academic image collections, including the Natural History Museum and the Wellcome Trust Medical Image Library.

While our original motivation was to create a data web of biological research images, this JISC-funded *Defining Image Access Project* has not been so limited, but rather has involved a broader exploration of image holdings in institutional repositories, without restriction to any particular subject area. Through separate projects, we are also pursuing data web activities related to online journals and to other digital image collections, and are investigating best practices for cooperative construction of domain-specific metadata schema (ontologies): these activities lie outside the scope of the *Defining Image Access Project*.

The original project proposal drew considerable inspiration from the JISC-commissioned report of Swan and Awre entitled *Linking UK Repositories: Technical and organizational models to support user-oriented services across institutional and other digital repositories* (Swan and Awre, 2006),

that was published shortly before submission of the *Defining Image Access* Project grant application in June 2006, and we are pleased to acknowledge our indebtedness to them.

7.2. Abstract from JISC *Defining Image Access* Project grant application

The *Defining Image Access Project* is a short requirements analysis project to investigate what would be required to develop and provide discovery and delivery interoperability for image data held in DSpace, EPrints and Fedora-based institutional repositories. The PI and his research officers will work closely with each of the Partner institutions, with JISC officers involved in the JISC e-Framework, Information Environment Architecture and Intute, and with the PIs of other JISC projects such as MIDESS, eBank, Preserv, CLIC and TASI. The work to be undertaken falls under four headings. **1 Repository structures:** To examine differences between the software structures and semantics of DSpace, EPrints and Fedora repository systems, as they affect image storage. **2 Image metadata:** To explore the suitability of existing repository metadata standards and schemas for handling image data, the granularity of existing metadata descriptors for images within repository holdings, and their semantic consistency across repositories; to determine how and where these metadata standards might need to be extended or improved; and to define the requirements for an image data web core ontology that will provide a basis for linking data across different repository sources. **3 Metadata harvesting:** To survey existing repository metadata harvesting and cross-searching systems, and additional Web tools for querying remote data sources, identifying candidates that best lend themselves to our goal of distributed image access. **4 Data web technology:** To undertake a systematic survey of lightweight software tools that will permit the creation of a central data web metadata registry, and to design the operational logic for data web functionality. Our work will include experiments using trial data, but the development of a fully fledged pilot / demonstrator will form the core of an anticipated larger follow-on JISC project. This project's deliverable will be a **Project Report** that will **(a) detail the findings and conclusions** from our investigations, **(b) recommend best practices** that should be supported by the JISC and adopted to enhance image interoperability between institutional repositories, **(c) provide implementation guidelines** for the creation of data webs, for use by those running institutional repositories, and **(d) identify existing open source software systems** that can provide elements of the desired data web functionality. We will also create a project Web site, a Wiki and an e-mail discussion list, and will hold two workshops meetings at which we can bring experts together for seminars and more informal exchanges. This project will contribute to existing JISC RDN / Intute harvest-based search services running over UK repository content, and to a larger vision of data webs that will provide interoperability between repositories, on-line journals, museum collections and other digital resources.

7.3. Project description

The JISC *Defining Image Access* Project has been a requirements analysis project of short duration to investigate the feasibility of establishing data webs that would permit cross-searchable integration of institutional repository image collections on a subject-by-subject basis using lightweight Semantic Web techniques.

In the *Defining Image Access* Project, we set ourselves the task of determining, for images within institutional repositories, what would be required to expose domain-specific image metadata in a manner whereby it could be indexed and made cross-searchable, with links back to the repositories for retrieval of the original images, thereby creating data webs for images relevant to particular knowledge domains.

Our goal was thus to define how best to enable integrated search operations across several repositories to find, for example, all photographs of the actor Sir Laurence Olivier performing in any Shakespearean play before 1970, or all images showing expression levels of the *aly* gene in the fruit fly *Drosophila melanogaster* at various anatomical loci.

7.4. Aims and objectives

Our *original* ultimate objective was to create image webs to integrate the search for published images of all types and subjects across heterogeneous institutional repositories. To do this requires mechanisms first to access and match metadata from different images sources, and then to locate those images that have associated metadata meeting particular criteria. We thus planned to undertake surveys of the repository image holdings of four leading universities located within convenient reach within the South of England, at Cambridge, Imperial College, Oxford and Southampton Universities (based variously on DSpace, Fedora and EPrints repository software systems, the three Open Source software systems commonly employed throughout UK universities), in terms of their software systems, holdings and metadata exposure mechanisms, hoping thereby to learn general lessons applicable to all UK university repositories. We also planned to survey software tools, services and metadata standards that would make the construction of such data webs feasible. On the basis of this information, we intended to develop strategies and software designs to create a data web design and an implementation plan to meet our original goal, which we could implement as a demonstration prototype within an anticipated JISC follow-up project. While we planned to undertake some limited piloting as part of our evaluation work, development of such a pilot system was never a goal to be implemented within this project.

While our objective has not changed, actual findings have led us to a significant revision of this initial plan, as explained in Section 15.1 below.

7.5. Approaches

Our basic approach was predicated on the notion that by using widely deployed web software components and techniques, we might short-circuit many of the development complexities that escalate the cost and limit the deployability of more complex information-sharing systems such as Virtual Research Environments. Our investigations coincided with the availability of several highly developed and work-hardened Semantic Web software tools, in addition to the many more conventional and widely used web server and content management systems. We aimed to focus our efforts on information design and software selection, rather than software design. We anticipated that our data web design would use Semantic Web techniques to combine information from to common data objects in heterogeneous repositories.

Some initial guiding principles for our approach were:

- To work with existing repository and their published metadata as we found them, thereby minimizing the technical impact that a data web system would have on the systems currently used by repository providers and their users (although we did, of course, aim to influence practice for the better!).
- To leave control and management of repository content and access firmly in the hands of the publishing institutions.
- To maintain full visibility of the existing repositories, leading users back to the original data sources, rather than acting as a portal through which they are accessed.
- To work as far as possible within the World Wide Web architectural framework, using the Web as an integrating platform.
- To design for use of existing open standards and open source software tools that could handle the 'heavy lifting' for an image data web implementation.
- To design around the use of lightweight web application technologies, with loose coupling between existing systems, thereby maximizing opportunity to replace or update any element of the technology used.

- To articulate criteria for the design a core metadata schema (ontology) for each data web, to which different repository metadata schemas could be mapped.
- To maintain consistency of our recommendations with the JISC's strategy for repository development and the JISC Information Environment architecture (http://www.jisc.ac.uk/whatwedo/themes/information_environment.aspx; <http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/>), while encouraging evolution of these to encompass Semantic Web technologies.

The information we have gathered, and our interpretations of it, have been lodged in our publicly accessible *Defining Image Access* Project wiki ([http://imageweb.zoo.ox.ac.uk/wiki/index.php/Defining Image Access](http://imageweb.zoo.ox.ac.uk/wiki/index.php/Defining_Image_Access)).

Alongside the primary approach noted, we experimented with some collaborative web-based tools for managing the project itself, and for gathering and publishing information. These included use of Semantic MediaWiki (http://meta.wikimedia.org/wiki/Semantic_MediaWiki) for the project wiki, use of the Drupal web content management system (<http://drupal.org>) for managing data files, presentations, etc., and use of WebCalendar (<http://www.k5n.us/webcalendar.php>). Use of these tools provided insights into possibilities for new web-based styles of collaborative working that informed the application of our data web ideas for repository access, and also influenced the style and content of this report.

7.6. Project deliverables

From the outset, the principal deliverable of the *Defining Image Access* Project was intended to be this Final Report that should (a) detail the findings and conclusions from our investigations, (b) recommend practices that should be supported by the JISC and adopted by institutional repositories to enhance image interoperability between them, (c) provide an implementation guideline for the creation of data webs, and (d) identify existing open source software systems that might provide elements of the desired data web functionality.

The project wiki, [http://imageweb.zoo.ox.ac.uk/wiki/index.php/Defining Image Access](http://imageweb.zoo.ox.ac.uk/wiki/index.php/Defining_Image_Access), intended initially just as an internal working repository for work in progress, has grown to become an information resource in its own right, of which third parties are making significant use and to which we intend to continue adding information as our understanding grows through work in follow-on projects. This constitutes an additional deliverable of the project.

We held four project workshops, which brought together both the project's consultant partners and selected external experts for seminars and more informal exchanges. The presentations from these workshops, with the authors' permissions, are available on our content management system Drupal at <http://imageweb.zoo.ox.ac.uk/drupal/>, forming a further deliverable.

A significant additional deliverable of this project has been the creation of a version of EPrints optimized as a repository for research images. Our experience here (<http://imageweb.zoo.ox.ac.uk/wiki/index.php/DefiningImageAccess/Tool/Eprints>) can be used by institutions wishing to install their own image repositories.

8. PROJECT ACTIVITIES

8.1. Project meetings

We organized and hosted four project meetings in Oxford: a kick-off meeting to solicit ideas and suggestions from project participants, two meetings to discuss tools and technologies and interactions with other JISC activities, and a final project meeting that was used to present our draft findings and solicit comments on our proposed future directions.

- Project Kick-off Meeting, 5 January 2007; ([Meetings/20070105/DefiningImageAccess-KickOff](#)).
- Second Project Meeting, 9 February 2007: *Tools and Technologies for Semantic Interoperability Across Scholarly Repositories*; ([Meetings/20070209/DefiningImageAccess-ToolsAndTechnologies](#)).
- Third Project Meeting, 9 March 2007: *JISC Interactions Meeting*; ([Meetings/20070309/JISC-Interactions](#)).
- Final Project Meeting, 22 June 2007: *Images and Repositories, the way forward*; ([Meetings/20070622/DefiningImageAccess-FinalMeeting](#)).

The main project page, at [Defining Image Access#Project meetings](#), contains links to agendas and notes for the various project-wide meetings, and to the speaker presentations.

8.2. Survey work

We undertook surveys of our consultant partners' institutional repositories, and of standards, software tools, services and related projects relevant to creation of image data webs. The initial list of topics to survey came from our own prior knowledge and from suggestions made by participants at the project kick-off meeting. Additional topics were added through the life of the project, as we became aware of them and their significance. An overview of the information we sought to obtain in the survey of related software projects, tools and standards is provided at the wiki schema page, [DefiningImageAccess/ReviewSchema](#).

The results of this survey work are recorded in the project wiki, linked to from the page [DefiningImageAccess/RelatedWork](#) (also accessible via the "Related work" link in the web page sidebar, under "defining image access"). In the section [Defining Image Access#Technical notes](#), there are also references to some other background information:

- [DefiningImageAccess/Resources](#) contains links to other resources that have not yet been actively surveyed.
- [DefiningImageAccess/Articles](#) contains links to further reports and articles.

8.3. Meetings with repository partners

We visited our project consultant partners involved in running institutional repositories, to learn about their repository systems, image collections, metadata usage and deployment, and other aspects of their operations.

- DSpace@Cambridge: [Meetings/20070122/Dspace@Cambridge](#) - meeting with Patricia Killiard and Tom De Mulder at Cambridge University Library.
- EPrints@Southampton: [Meetings/20070416/DefiningImageAccess-ECS-Southampton](#) - meetings with Jessie Hey and various EPrints repository colleagues at ECS, Southampton University.

- Repositories of the Oxford University Library Service: [Meetings/20070503/DefiningImageAccess-SERS-Oxford](#) - meeting with Sally Rumsey, Neil Jefferies and Alexander Huber of OULS/SERS to discuss the Oxford Research Archive and its image collections.
- Imperial College: [Meetings/20070615/DefiningImageAccess-ImperialCollege](#) - meeting with Dolores Iorizzo and colleagues to discuss the Imperial College image collections.

Links to notes from these meetings can be found at [Defining Image Access#Repository Meetings](#). The findings from our survey of repository systems, based on these meetings and also our own investigations, are recorded at [DefiningImageAccess/RepositorySurvey](#). The main part of the repository survey work was based on the Cambridge and Southampton collections, to which we had earliest access. The subsequent Oxford and Imperial College discussions served mainly to confirm the paucity or fragmentary nature of repository provision as far as image collections are concerned.

8.4. Software evaluation

The number of potentially useful software packages turned out to be many times greater than those upon which we could reasonably perform any meaningful hands-on evaluation. The choice of packages evaluated was based on what were perceived to be key components of an image web.

Our principal findings and conclusions are outlined in Section 10 below. For more details of the software evaluation we planned and conducted, see [DefiningImageAccess/SoftwareEvaluation](#).

8.5. Technical design activities

The original *Defining Image Access* project plan included construction of a technical design to implement a demonstration data web for image collections held by institutional repositories. Content alignment between heterogeneous data sources is known to be a difficult problem. It was our hypothesis that study of particular subject domains would reveal information design patterns that could be easily exploited for quick gains in coordinating and cross-referencing information from different sources. However, the lack of repository images and metadata uncovered by our survey work, described in Section 9.2 below, leaves this particular goal unrealizable at the present time.

Our parallel research activities within our ImageStore Project, described in Section 11.7 below, have confirmed that researchers typically give little thought to their long-term data preservation needs. Precious research datasets are commonly stored on the hard drives of individual researchers' computers, which may or may not be regularly backed up, or on uncatalogued physical media items (videotapes, CDs, DVDs, etc.). The originating investigators remain the *de facto* curators of these artefacts, because neither technical support nor funding is provided for curating and making them accessible, and also in some cases because of cultural resistance to data sharing and to new methods of information management. In particular, submission of publications and datasets to institutional repositories is not part of present-day research culture.

As a consequence of these findings, we chose to broaden the focus of our technical design efforts to include bridging the gap between current research practice and institutional policies for repository population, publication and preservation: that of capturing image metadata early in the research workflow, exposing it through publication in research databases and institutional repositories, and thereafter creating data webs between such published images and journal articles based upon them.

8.6. Dissemination activities

The following four presentations preceded this JISC *Defining Image Access* Project and were instrumental in its formulation:

Shotton, D.M. (2006) Data Webs: Web 2.0 alternatives to databases. Proc. *Semantic Interoperability for e-Research in the Sciences, Arts and Humanities* meeting, Imperial College, 30 March 2006. http://cidoc.ics.forth.gr/workshops/london_workshop/Shotton.pdf.

Shotton, D.M. (2006) BioImageWeb – integrating biological image data. Proc. 1st Research Information Network Workshop *Data Webs: new visions for research data on the Web*, Imperial College, 28 June 2006. http://www.rin.ac.uk/files/10_Shotton-BioImageWeb.ppt.

Shotton, D.M. (2006). Data Webs: new visions for research data on the Web. Given at a conference entitled *The Closed World of Databases meets the Open World of the Semantic Web*. National e-Science Centre, Edinburgh, 12 October 2006. <http://www.nesc.ac.uk/action/esi/download.cfm?index=3303>

Shotton, D.M. (2006). ImageWeb: a new vision for sharing published research images on the Web. Blackwell Publishing, 22 November 2006. <http://imageweb.zoo.ox.ac.uk/drupal/files/Shotton - ImageWeb - Blackwell presentation 22 Nov 2006.pdf>.

The following five presentations have been given as dissemination activities during the course of this JISC project:

Shotton, D.M. (2007). Image Semantics and Image Sharing. Symposium: The Digital Image. Oxford e-Research Centre and Department of Zoology, Oxford, organized by Dr Shotton on 16 March 2007. PowerPoint presentation available from our digital asset management system at <http://imageweb.zoo.ox.ac.uk/drupal/files/Shotton - Image Semantics and Image Sharing 16 March 2007.pdf>.

Shotton, D.M. and Klyne, G. (2007). Data webs for locating research images stored in heterogeneous distributed repositories. School of Electronics and Computer Science, University of Southampton, 16 April 2007. PowerPoint presentation available at <http://imageweb.zoo.ox.ac.uk/drupal/files/Shotton Southampton seminar 16-04-2007.ppt>.

Shotton, D.M. (2007). Image semantics. UK Electronic Information Group meeting: Image Management in Bio- and Environmental Sciences: New Directions. University of Manchester, 31 May 2007.

Shotton, D.M. (2007). Research images as first class publication objects. Wiley-Blackwell Executive Seminar. Royal Society, London, 1 June 2007. PowerPoint presentation at <http://imageweb.zoo.ox.ac.uk/drupal/files/Shotton - Data as first-class publication objects - Wiley-Blackwell presentation 1 June 2007.ppt>.

Shotton, D.M. (2007). Defining image access. JISC Digital Repositories Conference: Dealing with the Data Deluge. University of Manchester, 6 June 2007. PowerPoint presentation at <http://imageweb.zoo.ox.ac.uk/drupal/files/Shotton - Defining Image Access - JISC Repositories meeting 5-6 June 2007.ppt>, and from the JISC at http://www.jisc.ac.uk/media/documents/events/2007/06/david_shotton.ppt.

Others' presentations given at *Defining Image Access* Project meetings are accessible from Drupal:

- http://imageweb.zoo.ox.ac.uk/wiki/index.php/Defining_Image_Access#Presentations

On 16 March 2007, Dr Shotton organised a one-day Oxford e-Research Centre Symposium entitled *The Digital Image*. The programme and the presentations given at this symposium are available here:

- <http://www.oerc.ox.ac.uk/oerc/events/digital-image.xml>.

See also our wiki commentary on the symposium at:

- [Meetings/20070316/TheDigitalImage](#).

9. REPOSITORY SURVEY FINDINGS

In the following sections, the findings of our project are grouped together by subject, rather than chronologically by project activity undertaken, since this provides the reader with direct access to all the information on each topic under a single heading. These subjects are

- Evaluation of repository systems
- Evaluation of repository image holdings
- Evaluation of standards for repository metadata exposure
- Commentary on Web and Semantic Web standards
- Evaluation of software tools
- Survey of related R&D projects

9.1. Evaluation of repository systems

Our survey focused on just three repository software systems: [DSpace](#), [Fedora](#) and [EPrints](#).

9.1.1. DSpace

- <http://imageweb.zoo.ox.ac.uk/wiki/index.php/DefiningImageAccess/Tool/DSpace>.

The current version (v1) of DSpace software is a large monolithic program, which we understand has become difficult to maintain and adapt in its current form. The software is being restructured and/or rewritten for the next version, which is intended to be more modular and adaptable: it is too early to tell how successful this effort may be.

9.1.2. Fedora

- <http://imageweb.zoo.ox.ac.uk/wiki/index.php/DefiningImageAccess/Tool/Fedora>.

Fedora is generally agreed to be the most flexible of the repository systems, but this flexibility comes at the expense of not providing specific support for any particular mode of operation. Neil Jefferies of Oxford University Library Service, our consultant partner in charge of Fedora deployment at the University of Oxford, describes Fedora as an architectural product rather than just a software application; that is, a suitable element of a long-term preservation strategy, rather than a complete system to solve all today's problems. Fedora supports the concept of content models, but for this special handling (including metadata presentation) must be implemented, in the form of a content model implementation module that needs to be created and installed alongside the main repository management software. At Oxford, the Fedora repository core software is being deployed in association with the commercial software application VITAL (<http://www.vtls.com/Products/vital.shtml>), which provides front-end user deposit and access services.

9.1.3. EPrints

- <http://imageweb.zoo.ox.ac.uk/wiki/index.php/DefiningImageAccess/Tool/Eprints>.

Version 3 of the EPrints software has recently been released, and is claimed to have a very modular, extensible architecture compared with earlier versions. It is widely deployed and relatively easy to set up (for our own experience of initial installation of EPrints from scratch, see in Section 12.1 below). Customization consists in many cases of adding or editing Perl software scripts, or configuration files. However, although EPrints is relatively easy to install and customize, it does not provide much in the way of specific support for deposits other than conventional publications

(papers, theses, etc.), and it lacks in-built support for compound content objects, beyond multiple data files associated with a single deposit record.

9.1.4. Conclusions

Early survey results suggested that we would have little need to interact closely with repository software, but would be able to use OAI-PMH to harvest metadata from all repositories, so we made an early decision not to perform hands-on evaluation of the repository software systems themselves. Furthermore, it transpired that the actual deployment of these as institutional repositories was typically quite different, in their use of high-capacity back-end file stores and databases, from the kind of in-house deployment upon which we might base our evaluations.

Later in the project, a different role for repository software at the local research group level became apparent: for this we decided to evaluate just one system. Fedora was generally acknowledged to be an "architectural" component of a larger deployment, and as such unlikely to be easily deployed for a single research group. Because of the monolithic nature of DSpace Version 1, and its impending update to Version 2, we chose not to undertake detailed evaluation of DSpace v1. Rather, EPrints was chosen, with the advantage that it was already in use by the SERPENT Project (described in Section 9.2.5 below) to implement an image repository with provision of the kinds of domain-specific metadata handling that we required.

9.2. Evaluation of repository image holdings

- [DefiningImageAccess/RepositorySurvey](#).

We were surprised at the paucity of image collections in the institutional repositories we surveyed. At the institutional level, the dominant collection focus is on electronic forms of written papers (e-prints and e-theses), and we found the specialized support needed for collections of image and other datasets to be mostly absent.

9.2.1. Dspace@Cambridge

Of the four university repository systems we surveyed, the Dspace@Cambridge repository held the most images, in a small number of substantial image and video collections:

- **Anatomy**: partial outputs from the Department of Anatomy's Multi-Imaging Centre and Anatomy Visual Media Group.
- **Anthropological Ancestors**, including video interviews with famous anthropologists.
- **Archaeology**: Photographs from the Kilise Tepe project excavations, <http://www.dspace.cam.ac.uk/handle/1810/31289>.
- **Cambridge Rock Art Database**, <http://www.dspace.cam.ac.uk/handle/1810/61>.
- **Digital Himalaya, Digital Orient**, and **Nepal** materials.
- **Royal Commonwealth Society photograph project**, <http://www.dspace.cam.ac.uk/handle/1810/752>.
- **Scott Polar Research Institute** photographic collection (currently being digitized), <http://www.dspace.cam.ac.uk/handle/1810/183634>.
- **Social Anthropology**, <http://www.dspace.cam.ac.uk/handle/1810/23>.

However, the available metadata for these images is largely confined to the standard Dublin Core (DC) elements.

9.2.2. The Southampton EPrints repository

The Southampton EPrints repository is mainly devoted to print publications saved as PDF documents, within which many images lie buried, being neither described by metadata nor independently accessible. A small number of individual images are distributed unevenly throughout the holdings, particularly of works of fine art.

The number of repository holding having a particular file format can conveniently be determined using ROAR, the Registry of Open Access Repositories (<http://roar.eprints.org>). For example, <http://roar.eprints.org/?action=profile&url=http%3A%2F%2Feprints.soton.ac.uk> shows that there are 35 JPEG and 3 TIFF images within the Southampton University EPrints Repository.

9.2.3. Imperial College

Imperial College currently holds large amounts of scientific image data, held in personal, research group or departmental databases, including:

- **Centre for Bioinformatics Image Collection**, specializing in chemistry, genetics and biology, <http://www.bioinformatics.imperial.ac.uk/>.
- **Centre for Population Biology Image Collection**, <http://www3.imperial.ac.uk/lifesciences/research/nerccentreforpopulationbiology>.
- **Centre for Structural Biology Image Collection**, <http://www3.imperial.ac.uk/structuralbiology>.
- **Multimedia and Information Systems Image Collection**, images and videos of arts, humanities and science data, <http://mmis.doc.ic.ac.uk/index.html>.
- **The Science Museum Image Collection** of 17th-19th century scientific instruments is also available through Imperial College.

However, these are not presently part of an institutional repository, nor interoperable with other collections of similar type.

9.2.4. University of Oxford

As with the other institutions, most of the research and teaching images of Oxford University are held in independent academic collections such as OxCLIC (<http://wiki.oucs.ox.ac.uk/Itg-public/OxCLIC>). The Oxford University Library Service has recently established the Oxford Research Archive (ORA; <http://ora.ouls.ox.ac.uk/access/>). Initial emphasis has been on populating this with collection of published papers and e-theses, rather than other forms of data. This position is currently being re-thought, partly as a result of the forthcoming closure of the Arts and Humanities Data Service (<http://ahds.ac.uk/>), which has left many academics seeking new homes for their data. At present, however, ORA contains no image collections. A separate repository, the Oxford Digital Library (ODL; <http://www.odl.ox.ac.uk/>), contains a disparate collection of images, mostly of scanned historical scientific and humanities library holdings, which are detailed at <http://www.odl.ox.ac.uk/collections/index.html>.

9.2.5. Conclusions

Image collections exist in repositories, but there seems to be a lack of consistent and interoperable mechanisms to access them. Across the board, where image collections were available, we found that their metadata quality is variable: even generic metadata (i.e. basic Dublin Core) was not always consistently provided, and there was very little domain-specific metadata of the kind that would indicate the content of an image, or provide a context for its interpretation (e.g. for gene expression images, minimally the name of the expressed gene, and identification of the anatomical region and organism within which it is expressed).

From our discussions with institutional repository managers, we judged that current attention on papers and theses, which are well-served by Dublin Core metadata terms, had dominated

consideration of additional metadata requirements of images and other forms of ‘opaque’ data. This situation has recently changed: with the proposed closure of the Arts and Humanities Data Service in March 2008 (<http://ahds.ac.uk/exec/news/ahrc-news-may07.htm>), repository managers are now considering more carefully what they can do to accommodate data submissions.

The shining exception to these generalizations was the SERPENT repository ([DefiningImageAccess/Repository/SERPENT](http://archive.serpentproject.com), <http://archive.serpentproject.com/>), a specialized repository at Southampton University that contains an exemplary collection of carefully curated images and videos of sea creatures, incidentally captured by remotely operated cameras used to inspect submarine oil pipelines. Although this site uses EPrints software, it is operated as an independent specialized project repository rather than as an institutional repository, and some customization has been applied to enable it to handle domain-specific metadata. As an exemplar, SERPENT has had considerable influence on our thinking about repository-based collections of research images.

9.3. Evaluation of standards and protocols for repository metadata exposure

9.3.1. Metadata standards

Our survey of metadata standards included:

- Generic metadata standards:
 - **Dublin Core** (<http://dublincore.org/>),
 - **Qualified Dublin Core** (<http://www.dublincore.org/documents/2000/07/11/dcmes-qualifiers/>), and
 - the **JISC EPrints Scholarly Works Application Profile** for Dublin Core (http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile).
- Structural and object packaging standards:
 - **MPEG-21 DID** (MPEG-21 Digital Item Declaration, <http://www.chiariglione.org/mpeg/technologies/mp21-did/index.htm>), and
 - **METS** (Metadata Encoding and Transmission Standard, <http://www.loc.gov/standards/mets/>).
- Structural metadata relating to images and media files:
 - **EBU core** (European Broadcasting Union application profile; http://www.ebu.ch/CMSImages/en/tec_doc_t3293_tcm6-10494.pdf),
 - **VRA Core** (Visual Resources Association Core Metadata Element Set; <http://www.vraweb.org/>; <http://www.ukoln.ac.uk/bib-man/factfile/metadata/vra-core/>), and
 - **Z39.87**, a NISO standard that defines technical metadata for digital still images (http://www.niso.org/standards/standard_detail.cfm?std_id=731).
- Higher-level elements with potential for across-domain content applicability:
 - **CIDOC CRM Core** (http://cidoc.ics.forth.gr/working_editions_cidoc.html; <http://eprints.ecs.soton.ac.uk/12828/>),
 - **INDECS** (Interoperability of Data in e-Commerce Systems, <http://www.indecs.org/>),
 - **SKOS** (Simple Knowledge Organisation System; <http://www.w3.org/TR/swbp-skos-core-guide/>), and

- **FRBR** (Functional Requirements for Bibliographic Records), <http://www.frbr.org/>.
- Other more specialized metadata schemas:
 - **CCLRC Scientific Metadata Model** (<http://epubs.cclrc.ac.uk/bitstream/485/csmdm.version-2.pdf>),
 - **PREMIS data dictionary** for preservation metadata (<http://www.oclc.org/research/projects/pmwg/>), and
 - **SCORM learning objects** (<http://www.adlnet.gov/scorm/index.aspx>).

We fully expect to add more to this list over time, as they come to our attention. It is difficult to evaluate the various standards in isolation from specific applications, so at this stage we regard awareness of them as the main result of this section of our survey. Two deserve further comment:

9.3.2. Dublin Core

It is possible to capture a small amount of structured domain-specific metadata in the **dc:subject** field, but this is not really sufficient to capture sufficient context for the reliable interpretation of an image. The **dc:description** field can be used to convey free-text commentary on the content, which might be useful if used judiciously, while **dc:type** can provide additional clues and cues concerning the nature of the content. Similarly, **dc:coverage** can provide topic or spatial information about the content. But, even within a single collection, these fields are rarely used consistently, and they do not offer a practical basis for image discovery based on subject.

An attempt to regularize the use of Dublin Core has been made in the JISC Application Profile for Scholarly Works, also known as the EPrints application profile (http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile), that has been used to provide the metadata for this Final Report (Appendix B). The influence of FRBR in this is clear to see. FRBR itself goes in to far too much detail to be viable as a metadata standard in its own right: the information landscape that FRBR attempts to describe changes too fast for the vocabulary to keep up. (An example of this is that FRBR provides terms for describing the groove pitch and rotational speed of a gramophone record, but does not have any way to describe the bit rate of an MP3 data file.) We feel an important lesson to be taken from this is that common metadata schemes should try to describe the enduring properties of the domain, and leave more evolutionary matters to specialized vocabularies.

- See also:
http://imageweb.zoo.ox.ac.uk/wiki/index.php/DefiningImageAccess/RelatedWork#Metadata_standards_and_specifications

9.3.3. CIDOC CRM Core

The one metadata standard that has struck us as particularly relevant for providing a common structural framework for describing experimental data obtained from scientific observations, is the CIDOC CRM Core (<http://eprints.ecs.soton.ac.uk/12828/>), which is a subset of the full CIDOC Conceptual Reference Model developed to describe cultural heritage artefacts (http://cidoc.ics.forth.gr/working_editions_cidoc.html). It consists of about twenty metadata terms capable of capturing complex relationships between agents and entities through descriptions of mediating events. This seems to us to be a clean metadata model that is very flexible, while being only a little more complex than Dublin Core. Further, we understand that Dublin Core can be represented using CIDOC CRM terms through the introduction of existential entities (which, in RDF terms, would be represented as blank nodes or uniquely-minted URIs).

9.3.4. Domain-specific metadata schemas and ontologies

We intentionally did not survey any domain-specific metadata schemas, such as the Gene Ontology (<http://www.geneontology.org/>) and MIBBI (<http://mibbi.sourceforge.net/>) for bioinformatics,

although it is exactly this kind of domain-specific metadata that will be important for discovering and interpreting images. It appears that the provision of support by institutional repositories for specific and evolving domain-specific metadata is not a realistic prospect. We thus need to design alternative ways to handle domain information within the common repository metadata structures that are currently available, as discussed in the next sub-section.

9.3.5. OAI-PMH

OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting; <http://www.openarchives.org/OAI/openarchivesprotocol.html>) is a protocol developed by the Open Archives Initiative for harvesting the metadata descriptions of the records in an archive, so that services can be built using metadata from many archives. It has the advantage that, once a site's metadata has been harvested, only new and changed records need be harvested subsequently in order to ensure an up-to-date record of the evolving repository content.

Support for OAI-PMH is ubiquitous in the repository systems we looked at (indeed, some commentators suggest that OAI-PMH support is what defines a repository), but we have discovered that it is not a panacea for accessing domain-specific metadata:

- First, it is administratively difficult and costly to deploy OAI-PMH to access varieties of domain specific metadata in an institutional repository, since the system has to be separately customized for each metadata schema.
- Furthermore, OAI-PMH cannot perform discovery based on domain-specific metadata values. The repository community model seems to be to use a separate service like OAIster (<http://www.oaister.org/>) for such operations.

Clearly, adding to a repository data described by a new metadata schema should not require bespoke tailoring by the repository administrative staff. However, given modest initial support from repository staff, we believe it would be possible to modify existing repository software systems to handle adequate domain-specific metadata in ways that would not require substantial and ongoing case-by-case administrative overheads. For example, the Southampton EPrints developer Christopher Gutteridge has suggested that, for EPrints, it would be possible to add a couple of additional structural metadata fields to specify the location and the format of a file containing domain-specific metadata, while our Consultant Partner Neil Jefferies has told us that for Fedora, access to domain-specific metadata would be enabled by definition of an appropriate content model and provision of external tool support based on OAI-PMH for basic access.

10. SOFTWARE TOOLS EVALUATION

10.1. Commentary on Web and Semantic Web standards

10.1.1. Web standards

Today's World Wide Web is built on three groups of underlying standards and formats:

- For identification: URIs (Universal Resource Identifiers; <http://www.ietf.org/rfc/rfc3986.txt>).
- For data transfer: HTTP (Hypertext Transfer Protocol; <http://www.ietf.org/rfc/rfc2616.txt>).
- For representation: HTML (Hypertext Markup Language) for document representation mark-up for presentation in a Web browser (<http://en.wikipedia.org/wiki/HTML>) and XML Extensible Markup Language) for encoding data (<http://www.w3.org/XML/>).

These standards and formats are not, of themselves, sufficient to define the working World Wide Web in all its richness and diversity (thus Web documents and data are also represented in other

formats including CSS, Javascript, JPEG, PDF, PNG, RSS, Word and XHTML), but they do form a core that is common to a very high proportion of Web activity.

While other data transfer protocols and information formats can be and are being used without fundamentally changing the nature of the web, the one element that cannot reasonably be replaced is the use of URIs: a web without URIs would be essentially different from and non-interoperable with the Web we use today.

10.1.2. Semantic Web standards

The World Wide Web Consortium has supplemented the basic Web standards by a series of others designed to support Semantic Web activities, of which three are foundational:

- The Resource Description Framework (RDF, <http://www.w3.org/RDF/>), that has an abstract syntax (<http://www.w3.org/TR/rdf-concepts/>) that permits one to make simple logical statements describing the relationships between entities (in the form of subject-verb-object triples), that combine to form graphs of logically related statements, and that can be written out (serialized) in a variety of forms that a computer can process:
 - RDF/XML (<http://www.w3.org/TR/rdf-syntax-grammar/>),
 - TRiX (<http://www.hpl.hp.com/techreports/2004/HPL-2004-56.html>),
 - Notation3 (<http://www.w3.org/DesignIssues/Notation3>),
 - Atom (<http://www.ietf.org/rfc/rfc4287.txt>), and
 - GRDDL (<http://www.w3.org/2001/sw/grddl-wg/>).
- SPARQL, the W3C-recommended domain-neutral query language and protocol for accessing RDF data (<http://www.w3.org/TR/rdf-sparql-query/>)
- The Web Ontology Language OWL (<http://www.w3.org/TR/owl-features/>).

Our ImageWeb activity is particularly concerned with standards for the description of image metadata, and we see the Semantic Web as providing appropriate tools for this. Within the data web framework, we have chosen RDF as the underlying standard for representing such metadata within data webs. To quote Dan Brickley, RDF provides "*a strategy for principled decentralisation in a world where unanticipated data re-use, unanticipated data extensions, are valued.*" (Brickley, 2005). We also propose to use SPARQL for communicating and querying image metadata.

10.1.3. Comparison of SPARQL and OAI-PMH

When OAI-PMH is the widely deployed protocol for accessing repository data and metadata, why are we recommending a different one? SPARQL has two distinct advantages over OAI-PMH for our purpose of creating a data web to integrate heterogeneous data sources:

- First, SPARQL permits query selection by domain-specific metadata values (for example “?image depicts Red Tree Vole” would be a query for all entities, locally designated by ?image, that have a “depicts” property with the value “Red Tree Vole”), while OAI-PMH cannot do this.
- SPARQL results are based on RDF, which links to the wider Semantic Web of data.
- More importantly, SPARQL can be mapped to provide queries against arbitrary data sources.

Indeed, a great advantage of SPARQL is that, in addition to providing access to RDF data, it can also be used to access non-RDF data formats and interpret them as if they were in RDF. Such a process of providing SPARQL access to non-RDF data is described as providing a **SPARQL endpoint** on that data source (http://wiki.ontoworld.org/index.php/SPARQL_endpoint), and is a

key element in enabling Semantic Web technologies to access the vast amount of non-RDF ‘legacy’ data that exists. For example, the use of D2R Server (<http://sites.wiwiss.fu-berlin.de/suhl/bizer/d2r-server/>) provides access to information stored in relational databases, while SquirrelRDF (<http://jena.sourceforge.net/SquirrelRDF/>) can be used to access LDAP data.

See also:

- [DefiningImageAccess/Resource/SparqlEndpoints](#): Here we describe some proof-of-concept experiments in which SPARQL queries were used to combine information from two independent endpoints, for *Drosophila* Gene Expression Images and for the Gene Ontology (<http://esw.w3.org/topic/SparqlEndpoints>).

10.2. Tools for building data webs

- http://imageweb.zoo.ox.ac.uk/wiki/index.php/DefiningImageAccess/RelatedWork#Software_systems_and_tools

We surveyed some 35 software tools with a view to identifying components from which an image data web could be constructed. Of those, the following are particularly relevant to our purposes. This selection is not definitive or final, but represents our current judgements based on a range of technical and non-technical considerations. The selections are somewhat biased towards solutions that will help us get up-and-running as quickly as possible with a demonstration data web implementation, rather than those with the greatest potential for long-term scalability. Omission of other tools covered in our survey from the list does not mean they have been dismissed from consideration, but rather that no specific role for them has yet been identified.

10.2.1. Tools for annotation

- [DefiningImageAccess/Tool/Connotea](#), [DefiningImageAccess/Project/Dictate](#)

Connotea (URL) is a Web-based system to allow third party annotation and tagging of journal articles. The JISC DICTATE Project integrated Connotea with EPrints software. We are considering both Connotea and the Dictate implementation as possible routes for annotating images and for linking image data with other external resources – in the SERPENT repository, publications about *Mollusca* are linked by this method to images of the Piglet Squid. It is not obvious at this time how well the Dictate/Connotea tagging approach will play with more formal ontologically organized data. Interaction with the Rich Tags project (also at Southampton) should produce some interesting insights.

10.2.2. Tools for faceted semantic browsing

mSpace

- [DefiningImageAccess/Tool/mSpace](#).

We had hoped to evaluate mSpace (<http://www.mspace.fm/>), the faceted semantic browser for RDF data developed at the University of Southampton, as a potential means of quickly deploying a user interface for browsing metadata and images presented by a SPARQL endpoint or a data web. However, we are still awaiting the release of an Open Source version.

jSpace

- [DefiningImageAccess/Tool/JSpace](#).

jSpace (<http://clarkparsia.com/projects/code/jspace/>) is a new software product that has been inspired by mSpace, of which we are currently making a preliminary evaluation for the same purpose.

While such tools may not always represent the best possible user interface for presenting information about a given domain, they should provide a service that can be used for quickly exploring new image discovery options based on available image annotations.

10.2.3. Tools for creating SPARQL endpoints over OAI-PMH

- [DefiningImageAccess/Tool/Joseki](#)

Data webs aim to create pan-repository access based on Semantic Web standards. Central to our requirements is software that can be used to construct a generic SPARQL endpoint for OAI-PMH repositories. For this purpose, [Sesame](#) and [Joseki](#) were evaluated, both being systems that combine a deployable server with an underlying programmers' toolkit, and both being widely used and actively supported. We also evaluated Allegro and Virtuoso, both of which appear to be powerful candidate systems, but did not take them further mainly because we are not aware of any significant use of these in the open source Semantic Web community, although interest in Virtuoso seems to be growing. Either of these might be considered candidates in the future if we encounter a need for a higher performance SPARQL endpoint server.

In the end, we decided that Joseki (<http://www.joseki.org/>) and Jena (<http://jena.sourceforge.net/>) provide most of the tooling we need to create a SPARQL endpoint for an OAI-PMH repository. Joseki has a query engine architecture that has already been used to support distributed query, and it supports relational database storage of RDF graphs, while Jena provides a mechanism for updating a data set served by Joseki. Further details are given in Section 15.5 below.

10.2.4. Tools for creating core data web functionality

The main functions of a data web will be (a) to accept a query using terms from a common schema and break it into sub-queries that can be presented against data from different sources, (b) map each sub-query into terms used by the corresponding data source, (c) submit these sub-queries against the various data sources using SPARQL, and (d) to collect the results and map them back into the common schema for presentation to the user.

The three main components that will provide this functionality (Figure 1) are:

- The Data Web Query Handling Service, which will incorporate the functionality of a SPARQL distributed query handler. The work done at HP Labs on distributed queries ([DefiningImageAccess/Tool/DARQ](#)) is highly relevant to our task of constructing the data web query handling service.
- The Data Web Schema Registry, which provides information used to map schema terms between the core schema and the source-specific schema.
- The Data Web Co-reference Service, which recognizes common entity references in different data sources possibly expressed using different identifiers.

The schema registry and co-reference services will operate through open-ended data source 'subscriptions' that supply the information required used to map their contents to a common form. Initially, we will focus on simple one-to-one term mappings and relationships, e.g. Red tree voles and *Arborimus longicaudus* are the same species, or the current UK Prime Minister is Gordon Brown. We will then introduce more complex mapping relations as and when they are useful, being guided in such complex mappings by the work of Kondylakis, Doerr and Plexousakis on schema mapping frameworks (Kondylakis *et al.*, 2006). We anticipate that the Schema Mapping Service will be more pattern-oriented, while the Co-reference Service will be based more on direct lookup of identifiers.

In all of these areas, general solutions are hard to devise. Our goal is not to create such a generalized system, but rather to examine the available data and incrementally build systems that are effective on the structures we find. We fully expect our systems to utilize specific knowledge of

the data sources, and will aim to encode that knowledge in ways that allows the software components to be re-targeted to other areas. This corresponds to the rule-based architectures described in the work of Doan and Halevy on relational schema alignment (Doan and Halevy, 2005).

At this time, our proposed implementation strategy for building the schema registry and co-reference service elements of a data web is as a new development based on a lightweight Web application framework such as Python Turbogears (<http://www.python.org/>, <http://www.turbogears.org/>), or Ruby on Rails (<http://www.ruby-lang.org/>, <http://www.rubyonrails.org/>).

11. SURVEY OF RELATED R&D PROJECTS

We surveyed a substantial number of new and existing projects related to our image data web vision, the majority of which have been funded by the JISC.

Brief descriptions of the JISC projects we surveyed can be found in our Semantic Media Wiki at <http://imageweb.zoo.ox.ac.uk/wiki/index.php/Special:SearchTriple?title=Special%3ASearchTriple&subject=&relation=&object=&attribute=JISCProject&value=True&do=Search+Attributes> (Note that this links to a dynamic search page, and the results may change as new information is added to the wiki). Full details of our survey results are given at http://imageweb.zoo.ox.ac.uk/wiki/index.php/DefiningImageAccess/RelatedWork#Other_projects_and_groups.

The following are brief commentaries on the relevance of each project to our image data web vision, arranged alphabetically by project name and with direct links to the relevant sections of the project wiki. Most of these project are also mentioned within the independent perspective on our *Defining Image Access Project* given by the UKOLN Repositories Research Officer Julie Allinson, that was written independently from this report and is included as Appendix A.

11.1. CAIRO

- [DefiningImageAccess/Project/CAIRO](#).

This project is focused on repository user interfaces for interacting with complex repository content. No specific contribution from this work to our plans is perceived at this stage, although there could be interactions with ingest mechanisms that we discuss elsewhere.

11.2. CLADDIER

- [DefiningImageAccess/Project/CLADDIER](#).

This project has many goals similar to our own long terms goals, notably the linking of research papers to raw research datasets. Because of technical differences in approach, tools from the CLADDIER project may not be directly helpful to us, but their experiences gained when working with atmospheric science data are clearly of direct relevance, particularly the integrated CLADDIER Discovery Service (<http://isegserv.itd.rl.ac.uk/claddier/search/single/>) for searching simultaneously across data stores and publication repositories, and the CLADDIER ‘ping’, a lightweight peer-to-peer protocol whereby data sources can communicate behind the scenes to establish links from cited to citing (Matthews *et al.*, 2007). Currently, not all the project outcomes are visible, but hopefully they will become so in time to influence our future work.

11.3. Common Repository Interface Working Group (CRIG)

- <http://www.ukoln.ac.uk/repositories/digirep/index/CRIG>
- <http://www.ukoln.ac.uk/projects/iemsr/>

We have had fruitful discussions with the two co-chairs of this JISC working group, particularly about the desirability of establishing SPARQL endpoints on institutional repositories, and the use of SPARQL as a common mechanism for accessing repository metadata, as a supplement to OAI-PMH. Other areas of common interest with this group include OAI-ORE, SWORD (discussed in Section 11.13 below) and their use for bulk ingest.

11.4. DExT

- [DefiningImageAccess/Project/DExT](http://www.ukoln.ac.uk/DefiningImageAccess/Project/DExT).

The DExT Project is developing data exchange tools based on XML and RDF Schema (RDFS; <http://www.w3.org/TR/rdf-schema/>). We anticipate that tooling produced by this project that could be useful in the creation of data webs.

11.5. Dictate

- [DefiningImageAccess/Project/Dictate](http://www.ukoln.ac.uk/DefiningImageAccess/Project/Dictate).

The Dictate Project has created a Connotea plug-in for EPrints software, to permit user annotations of individual repository holdings through the EPrints Web interface. Having concluded that we wish to enable post-deposit collection of image metadata, and also that we intend to use EPrints as the basis for our own research image publication repository, it is natural that we should also explore the role of Dictate to allow additional image annotations from third parties to be captured.

11.6. eBank-UK, R4L, SPECTRa

- [DefiningImageAccess/Project/eBank](http://www.ukoln.ac.uk/DefiningImageAccess/Project/eBank).
- <http://www.ukoln.ac.uk/projects/ebank-uk/>.
- <http://r4l.eprints.org/>.
- http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories/project_spectra.aspx.
- <http://www.lib.cam.ac.uk/spectra/>.

eBank-UK has been a seminal project in the area of linking research data to publications. We are hoping to draw advice from participants in this project in our own future work. R4L (Repository for the Laboratory) and SPECTRa are continuations of the eBank-UK work.

This early work was grounded in the field of chemistry and chemical crystallography, which is characterized by widely used and well structured data formats for many aspects of the work. Part of our challenge is to see if we can apply the same ideas to the more heterogeneous and loosely structured data generated by life science research, particularly where recorded images represent a fundamental element of the scientific record.

11.7. ImageStore project

- http://www.dcc.ac.uk/scarp/#image_store.
- [DefiningImageAccess/Project/SCARP](#).
- http://imageweb.zoo.ox.ac.uk/wiki/index.php/The_ImageStore_Project.

Our own ImageStore project, a sub-project within the Digital Curation Centre SCARP Project, is being conducted in collaboration with the e-Science Group of the Science and Technology Facilities Council at the Rutherford Laboratory, Harwell. We have been working with active biological research groups in Oxford to determine their requirements for the long-term preservation and re-use of biological image and video data. We expect that project to raise new ideas that will inform our ongoing work on image webs for biological research datasets.

11.8. Information Environment Metadata Schema Registry (IEMSR)

- http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres/shared_services/project_mregistry.aspx
- <http://www.ukoln.ac.uk/projects/iemsr/>

Our data web architecture includes a schema registry and a co-reference service, which will grow through a mechanism whereby new data sources are "subscribed" to a particular data web. The IEMSR notion of a central JISC metadata schema registry suggests some overlap in purpose, though the expected circumstances of use are very different. One possible area of cooperation that we have identified is the use of IEMSR to hold schema alignment rules for commonly used schemas.

11.9. Intute Repository Search

- <http://irs.ukoln.ac.uk/>

There is clearly some overlap of purpose between our vision for data webs across repositories and the proposed Intute Repository Search, though the approaches proposed are quite different. We can see a number of ways in which our services might usefully interact (e.g. data webs providing a programmatic interface to information, enriching the scope of the Intute Repository Search), and we are in ongoing discussions with the Intute team.

11.10. OAI-ORE

- [DefiningImageAccess/Project/ORE](#)
- [DefiningImageAccess/Project/Pathways](#)

The current Open Access Initiative's Object Re-use and Exchange (OAI-ORE) project is a multinational project funded by the Mellon Foundation and led by well known US experts in the repositories field, with significant involvement from JISC and UKOLN. It has grown from Pathways project. The ability to re-use and exchange repository data and metadata is central to our own goals, and it seems clear that we should track a project of this significance.

A recent white paper from the project (Lagoze and Van de Sompel, 2007) proposes use of Semantic Web technologies, specifically RDF named graphs (Carroll *et al.*, 2004), to describe compound objects. This approach will integrate well with our plans to capture RDF domain-specific metadata about the content of images.

11.11. Rich Tags

- [DefiningImageAccess/Project/RichTags](#)
- <http://www.mspace.fm/projects/richtags/>

A recurring theme on our work is how to move from researcher's informal descriptions of their observations to more formally codified records of those observations that are amenable to computer processing. The RichTags Project (<http://www.mspace.fm/projects/richtags/>) is one project of a few we have encountered that seems to be addressing this particular issue, and we are eager to incorporate any findings or tooling from this project in the user annotation systems we develop.

11.12. StORe

- [DefiningImageAccess/Project/StORe](#).

This project, like CLADDIER, has many goals similar to our own long terms goals, notably the linking of research papers to raw research datasets. In addition to working on a prototype system based on Web 2.0-style tools that include access control to allow managed dissemination of material, the StORe project has conducted a detailed survey into the nature of scientific research and publication, and the attitudes of researchers to institutional repositories, and has examined some the implications of including data publication in this process, validating many of the conclusions we have independently reached. The executive summary of the survey phase of the report (<http://jiscstore.jot.com/SurveyPhase>) makes for interesting reading concerning the researchers' attitudes to repositories:

StORe leader Graham Pryor (Pryor, 2006) says:

“Cultural and organisational barriers prevail in all disciplines, which serve to deter the deposit of research data in repositories, and an inherent culture of self-sufficiency in the generation and organisation of data militates against what might be viewed as prescriptive intervention by knowledge management professionals”.

Later in the same paper, when exploring working practices and concerns of practicing researchers, he suggests "the results from the StORe survey do imply that a step change is necessary".

While the findings resonate with what our researchers tell us, we question whether what is needed is a “step change”. Rather, we believe that we need to build on existing research practices and take account of researchers' concerns, building trust and showing how use of repository tools can augment rather than distract from or disrupt their research activities.

11.13. SWORD

- [DefiningImageAccess/Project/SWORD](#)

SWORD uses Atom and Atom Publication Protocol to support repository ingest. This approach sits very well with our philosophy of using lightweight Web protocols to support loosely coupled components. In particular, SWORD may provide the ideal mechanisms for moving images and metadata from local research group storage to a public-facing institutional repository.

12. PROJECT SOFTWARE DEVELOPMENTS

12.1. Creation of an EPrints repository for capturing research images and metadata

- [DefiningImageAccess/Tool/Eprints](#)

In order to obtain practical experience of building and publishing image collections in a repository using domain-specific metadata, we are conducting an experiment of setting up an image repository using EPrints.

To this end, we are adapting an instance of the EPrints 3.0 software system for use as a research group image publication platform. In so doing we hope to solve real problems being faced by our research community, while at the same time gaining an understanding of what is involved in creating an image repository enhanced with domain-specific metadata.

EPrints 3.0 was chosen as our experiment for the following reasons:

- (a) It is one of the well-established software systems in the digital library community for archiving digital items, including theses, reports and journal publications.
- (b) It has built-in support for the OAI-PMH protocol, which will allow the contents of our image repository to be harvested by other parties.
- (c) It has a built-in user interface, which makes it fairly quick to set up the repository and present it to users; and finally:
- (d) It has previously been adapted by the Southampton SERPENT repository (<http://archive.serpentproject.com/>) to publish images using domain-specific metadata, as described in Section 9.2.5 above.

Because the EPrints system is targeted at digital text archives, it has good support for the use of Dublin Core metadata for describing and searching for digital items. However, in order to use EPrints to store and publish our images with domain-specific metadata, we need to make the following changes:

- (a) Customization of the underlying database itself, in order to store domain-specific metadata along with images.
- (b) Customization of the user interface, in order to permit searching and browsing for images using domain-specific metadata, and displaying the metadata within the search results in the most useful manner.

The initial database customization itself has been realized in three steps:

- (a) Extending the database schema with the extra domain-specific metadata fields.
- (b) Augmenting the EPrints web form ingest process with a tool that permitted uploading of groups of images and of their metadata from available Excel spreadsheet files (i.e. performed bulk ingest).
- (c) Exposing the domain-specific metadata stored in the repository through the OAI-PMH protocol, along with the standard DC metadata.

The user interface customization has involved reordering the displayed metadata fields. Many of the standard DC metadata fields presented by the native EPrints' user interface are superfluous to our users' requirements, and have thus either been put into a secondary position or removed entirely, in order to let users concentrate on the domain-specific image metadata.

The initial installation and customization of the EPrints system, and extending it to host a substantial proportion of our local research team's collected images and metadata, has been a non-trivial undertaking, involving some six weeks of developer effort from a 'standing start'. The early results were demonstrated at the *Defining Image Access* final project meeting.

This hands-on experience has shown that:

- (a) it is feasible but not trivial to adapt the existing EPrints digital archive software system for the storage of images with extensive domain-specific metadata;
- (b) EPrints allows us to publish domain-specific metadata to Web pages viewed by end users and via the OAI-PMH protocol for machine processing; and finally
- (c) the built-in EPrints user interface support can be customized without too much difficulty to provide new functionalities.

The images being used for this experiment are *in situ* gene expression images recently acquired by researchers in our department to study the factors causing sterility in the fruit fly *Drosophila melanogaster*. Currently, thousands of images, accumulated during the last year or so of research work, are kept in a file system organized by the date they were created and described using an Excel Spreadsheet. This system makes it extremely difficult for researchers to locate images from the file directories by their domain-specific metadata, such as the name of the gene whose expression pattern is being determined, or the mutation whose effects are captured in the image. A proper image repository is needed to assist the researchers in uploading, storing, searching and publishing images with the domain-specific metadata.

The initial results of this repository implementation have been presented to our biological researchers to gather feedback and additional requirements. Their comments were positive, but further user interface customization was deemed desirable, such as highlighting the domain-specific metadata and presenting the images in a more user-convenient way. Also, researchers raised the need to integrate their own images held in this image repository with images from external repositories such as the Berkeley *Drosophila* Genome Project (<http://www.fruitfly.org/>) – in other words, **our research users want us to create an image web for *Drosophila* gene expression images!** Implementation of this suggestion will form part of our proposed future work.

This approach of using EPrints promised two key advantages:

- (a) by building on an existing technology platform, we can rapidly create a local publication tool for our research clients, and
- (b) by using an existing repository platform, we anticipate easing the transition to long-term storage in an institutional repository.

The first advantage has been clearly proven. While the initial implementation leaves much to be desired, we believe subsequent progress to improve it will prove quite rapid, thanks in large part to working with standard web technologies and simple interfaces. Concerning the purported advantage (b), we have no evidence at present, but are looking to the ongoing work of the OAI-ORE and SWORD projects to facilitate such automated migration between repository systems.

This notion, of local research data storage and subsequent ingest to an institutional repository, has become a key element of our proposed technical strategy: to collect and publish data locally by enhancing researchers' existing workflows, then to look to migrate selected collections to a more formally controlled repository environment. This is further described in Section 15 below. One advantage of this approach is that metadata can be maintained locally in its original richer form, with conversion or 'dumbing down' to more widely used metadata vocabularies, if required, at the time of migration. This avoids the potential danger of limiting initial metadata collection to conform to an expected external standard requirement that is generally weaker than what might be agreed and implemented for in-house use, and that may also change over time (e.g. from Dublin Core to the Scholarly Work Application Profile).

13. AREAS REQUIRING FURTHER EVALUATION

To achieve the goals we set ourselves for image data webs based on institutional repositories, we need to address some further issues of metadata capture and/or creation. The following areas have come to our attention:

13.1. Metadata acquisition

Useful metadata is often collected automatically during the process of image creation – automated data recording by digital cameras is a simple example of this. However, researchers may be reluctant to make their research data public at the time of first creation, choosing to wait until after their results have been published in peer-reviewed journals. Later, when papers describing the research have appeared, such metadata may have been lost, or the effort required to harvest it from the original media might be seen to exceed any potential benefit to the researcher. This suggests that the acquisition of automatically captured metadata is an issue that needs to be addressed separately from, and well in advance of, paper publication or formal data deposition in a repository.

13.2. User interfaces

All repository software systems permit images to be treated like other forms of data, leading to the temptation to think that images are no different from other forms of data. However, the visual nature of images means that user interfaces need to be specially designed to take into account the ways in which people like to work with images. Conventional search interfaces are not good for discovering images. Conversely, visually-oriented browse interfaces can be very helpful when trying to locate particular images in a collection. This is an area that needs special consideration when designing or selecting tooling for image collection repositories.

13.3. Ontologies and annotation

While ontologies are excellent for providing pre-defined controlled vocabularies for image metadata creation, it is often very difficult for researchers to agree in advance exactly what metadata information is needed. We need to seek out approaches that allow consensus to be codified, but yet do not prohibit recording of additional informal annotations or tags. The interaction between such formal and informal descriptive approaches in an active area for research and discussion, and indeed formed the topic for the Second Meeting of the Ontogenesis Network (<http://ontogenesis.ontonet.org/>). Peter Mika has made important contributions to this area (Mika, 2005), while within the UK the RichTags project (Section 11.11) and EnTag project (Appendix A, Section 6) are most interesting lines of investigation.

13.4. Schema alignment and co-referencing

We perceive advantages in separating schema-level alignment of different data web resources to a data web core schema to achieve semantic interoperability, which requires initial hand-crafting at the time of subscription to the data web, and subsequent instance-level alignment of content, which can then be automated (Zhao, 2007). However, we realise that this distinction between schema alignment and instance alignment is not always clear cut, and that there are some ambiguities in our approach.

We are aware of a fair amount of work on schema alignment applied to relational databases (Doan and Halevy, 2005; Franklin *et al.*, 2005; Halevy *et al.*, 2006; Kondylakis *et al.*, 2006), but less involving practical efforts to apply these ideas to loosely structured data. We feel there are important lessons to be learned by attempting a bottom-up implementation of these ideas in the

context of biomedical research data, starting with obvious specific alignments and working towards more general patterns.

With relational databases, the distinction between schema and table data is clear, as discussed by (Zhao, 2007). However, for Semantic Web data, the distinction is sometimes less clear, as there can be some lack of clarity about whether a term should be treated a class (i.e. an ontological or schema element) or as an instance (see, for example, <http://www.w3.org/TR/swbp-classes-as-values/>). Let us illustrate this with some examples:

- Animal and plant species are recognized as classes of objects, while Dolly the Sheep is an instance of a sheep. Here there is no ambiguity.
- Hydrogen is the class of hydrogen atoms, but statements about ‘Hydrogen’ might be treating hydrogen as an instance of the class ‘Elements’.
- *Aly* is the name of a genetic strain of the fruit fly *Drosophila*, a subclass of all *Drosophila*, but gene expression observations made on individual flies from this class, i.e. on instances, may be discussed as observations of *Aly*.

How do we handle such potential ambiguities? We propose to adopt a common sense approach: if a term appears in instance data, then recognizing different terms meaning the same thing in different instance data is a co-reference problem; however, if different data stores describe the same attributes using different schematic structures, then schema alignment is needed.

13.5. Metadata pre-harvesting versus distributed querying

The degree to which it will be useful for the data web to pre-harvest core metadata into a central metadata registry, rather than relying solely on distributed SPARQL queries, is a subject that requires practical research. PictureAustralia is an example of a service in which all metadata is pre-harvested and indexed. Experts with extensive experience of distributed database querying have warned us of the difficulties and performance costs of a system that relies fully on distributed query processing. Others have shared with us their experiences of poor performance when running queries over large unified RDF triple stores. We remain hopeful that distribution of SPARQL queries over distributed data resources equipped with SPARQL endpoints will indeed be feasible and show reasonable performance, but the proof of this will only come when we implement our proposed demonstrator data web (see Section 15 below). We reserve the option to allow a data web to perform selective harvesting or caching of metadata in response to possible performance constraints.

14. CRITIQUE OF THE PROJECT

14.1. Summary of project achievements

These have previously been presented in the Executive Summary (Section 2.2, q.v.), and will not be repeated here.

14.2. Project shortcomings

With more time and manpower, we could have undertaken more thorough survey work, and become more familiar with a greater variety of other developments in this area of data integration and knowledge management. Although this was never one of our goals within this project, it should be realized that we have not yet created a demonstrator data web, required to fully validate our ideas.

15. PROPOSAL FOR FUTURE WORK

15.1. A design outline for an image data web as part of an integrated workflow for research image data publication and reuse

We have written a detailed [presentation](#) entitled “[Towards a Technical Infrastructure for Research Image Data Publication](#)”, which gives a description of our vision of how a data web can fit within an integrated workflow for research image data publication, repository deposition and reuse. The following summary figure (Figure 3) is taken from that presentation, where it is explained.

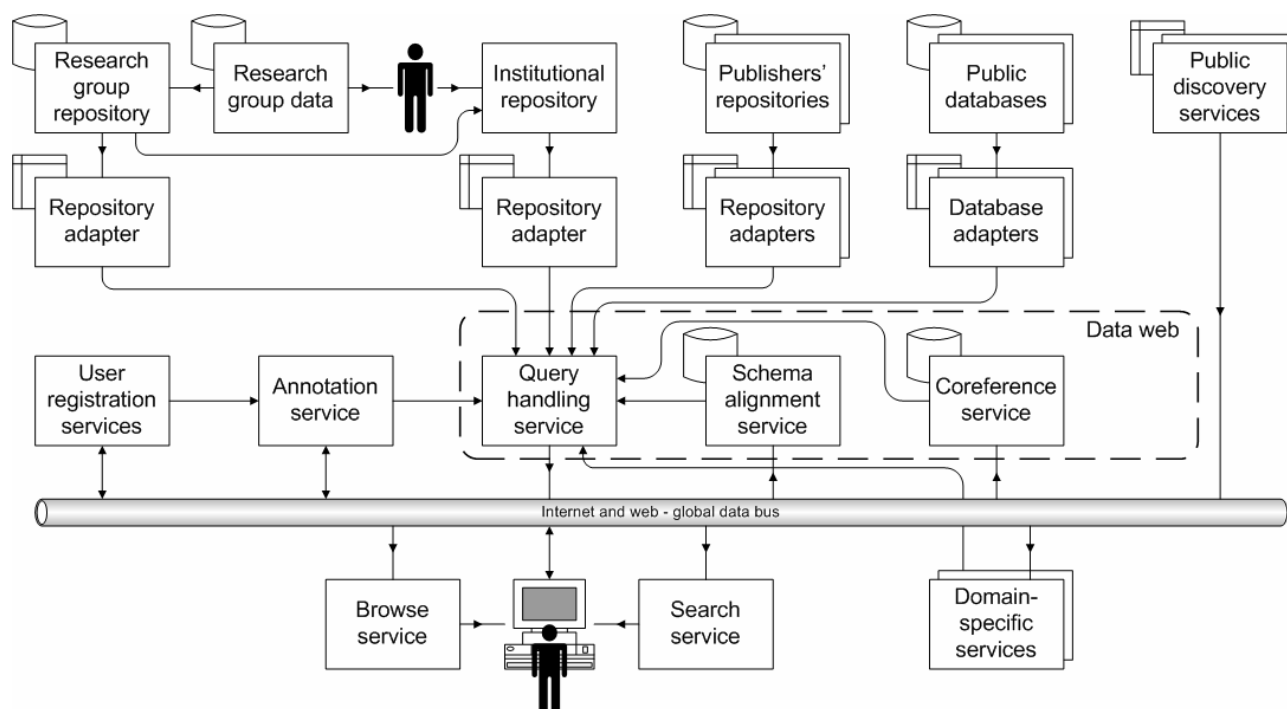


Figure 3. Diagram showing a data web within a wider research data publication environment

As mentioned in Sections 7.4 and 9.2.5 above, our original goal of creating subject-specific data webs between image holdings in institutional repositories cannot be fulfilled immediately. Our present goal is thus to develop a supportive software environment for the semantic enrichment of research image data as part of an integrated workflow, leading to their publication for inspection and reuse, and their subsequent deposition in institutional repositories for preservation. Within this environment, image data webs will provide integration of distributed research images and research publications wherever they occur, within institutional repositories, research group databases or elsewhere.

To support this, appropriate domain-specific metadata for research images needs to be created at the time of image capture within the laboratory or field research environments, in ways that enhance the individual researchers' activities and impose as little as possible by way of additional workload and cognitive overhead. This echoes a recommendation by Andy Powell at the recent JISC conference (Powell, 2007), and follows the 'annotations at source' method of the CombeChem Project (www.combechem.org/) (Hughes *et al.*, 2004). Images and associated metadata would then be placed in a local database such as our EPrints repository, and subsequently, if and when the researcher chooses, migrated to the institutional repository.

We further plan to add an annotation service that enables *post hoc* annotation of existing image collections, as a means of enabling third-party comments and observations to be recorded in a manner that does not violate the integrity of the content of the original image repository.

15.2. Software approaches to achieving a service-oriented architecture

There are two distinct (but not mutually exclusive) approaches to building Web-based data-sharing systems and distributed applications:

- Those based on a Web Services protocol stack consisting of SOAP, WSDL and sometimes UDDI, that utilize what is essentially a Web-based remote procedure call (RPC) framework. The Wikipedia article on Web Services (http://en.wikipedia.org/wiki/Web_service) provides a useful background to these.
- Those using a REST approach ('REST' stands for Representational State Transfer, a term introduced by Roy Fielding in his PhD Thesis (Fielding, 2000)), in which a small number of primitive operations are used to access and manipulate representations of the state of resources. While it is possible to implement a REST style using a SOAP-based Web Service stack, REST is more commonly associated with the use of unadorned HTTP to access and manipulate resources.

The term Service Oriented Architecture (SOA) is sometimes mentioned in connection with Web Services and REST. It is difficult to be definitive about SOA, as there are several extant definitions (see, for example, http://en.wikipedia.org/wiki/Service_Oriented_Architecture), but a key element is a focus on resource sharing between administrative domains by interacting with defined services, hiding implementation details. As such, SOA can be implemented either as a Web Service Remote Procedure Call (RPC) style, or using a REST style.

For creating data webs, we are particularly concerned with the ease of independent development and evolution of the software components, this being key to accessing resources that already exist on the Web. For this, the REST approach offers a key advantage by investing all of the application-specific semantics in data formats rather than in protocol specifications. The basic operations in a REST framework are common to all resources (commonly: http GET, PUT, POST, DELETE; or CREATE, READ, UPDATE, DELETE), and hence are domain independent. All domain-specific semantics, therefore, must be conveyed in the data that is exchanged. A data format can be constructed or interpreted by one application at a time, and some simple rules (e.g. "Ignore any parts that are not recognized") allow the providing and consuming applications to evolve independently. In contrast, protocol operations such as RPC involve two (or more) communicating parties, and require a degree of cooperation to handle changes to the protocol operations; this can mean that the communicating components must be updated in lock-step, making evolution of deployed functionality more difficult.

- See also http://imageweb.zoo.ox.ac.uk/wiki/index.php/Defining_Image_Access#Architectural_design.

15.3. Data web software framework

During the course of this project, our ideas for the structure of data webs have matured, gaining a clearer focus on the use of SPARQL as the RDF query language that can unify the distributed heterogeneous data resources contributing to a single data web. We have also come to the view that, rather than a data web just harvesting metadata conforming to a particular namespace from the web as a whole (a model that opens data webs to the potential of spamming), contributing data resources may 'subscribe' to particular data webs. To avoid an undue workload falling on the data web provider, it is envisaged that these subscription activities may be undertaken by each data resource individually, following guidelines and employing a core schema established for the particular data web by the data web provider.

We propose a software framework for creating domain-specific data webs, based on the fundamental components previously shown in Figure 1 (Section 6.3), that:

- avoids requiring content providers to change their source data,
- separates syntactic integration from semantic integration of data sources,

- allows a data web to grow through decentralized subscription of data sources,
- provides programmatic query access over the combined data, and
- provides a basis for a range of supporting tools for browsing, searching, further annotation, personalized access and more topic-specific services.

This is achieved through combination of the following elements:

- SPARQL endpoints for each data source, presenting all data in terms of the RDF abstract data model, but using schemas derived directly from each source. These will ideally be deployed with or near to each data source, but the framework does not require this;
- a core schema (or ontology) and entity reference schemes that are defined when a data web is established;
- hand-crafted semantic alignment rules that map data-source-specific terms to expressions in the core schema;
- a subscription mechanism (presenting semantic alignment rules) that incorporates new data resources into a data web (which can in principle be employed by administrators of the data web, a data source or any third party);
- a Schema Alignment Service registry that serves information about schemas used by subscribed data sources, used for alignment of their diverse data;
- a Co-reference Service that uses information about subscribed data sources to allow different identifiers to be recognised as referring to the same thing (such as different catalogue numbers for a particular artefact); and
- a distributed SPARQL Query Handling Service that fields incoming requests expressed using the core schema, presents sub-queries to the data-source SPARQL endpoints using their specific schemas, and collects, translates and combines the responses into a final result.

15.4. Implementation plan

Our technical developments, and the set of services that we will create, will share the following characteristics:

- Use of the Web as the platform.
- User-led design, with user evaluation and feedback throughout cycles of iterative development.
- Use of ‘agile’ test-led development techniques, developing simple testable prototypes early and expanding their functionality incrementally by iterative improvement.
- Loose coupling of services with programmatic access, in the spirit of the JISC IE Architecture.
- Use of third-party Open Source software tools and adoption of international standards wherever possible.

This data web framework is envisaged as a collection of loosely coupled software components that can, for the most part, be developed, deployed, evolved and evaluated independently of each other, providing resilience and ease of service evolution. Within this framework the common elements for interoperability are the use of RDF as the common data model and the use of SPARQL to access such data. We also envisage that some sources may be presented using a syndication format such as RSS and/or Atom (which can be used as a carrier syntax for RDF data), and which would enable linkage to mashup services such as Yahoo Pipes (<http://pipes.yahoo.com/pipes/>).

Since we do not propose that this framework be implemented as a single monolithic entity, we avoid the risks inherent in creating large and complex software systems that stand or fall on the successful integration of all components. A major problem with complex monolithic web-based systems such as some Web portals is their fragility, particularly when they are to be deployed in diverse environments or as parts of subsuming software systems (Klyne, 2006).

Instead, we imagine that specific user-centred projects will be identified that can use two or more of these components in combination, enabling a component-wise development of the framework driven by specific use-cases. This will allow the component implementations to be responsive to actual needs, and, in the style of agile development, will avoid complications of meeting hypothetical specifications that are never actually required. The component-based approach also makes it easier to incorporate existing tools without modification, either directly or with some lightweight shimming of their interfaces. A further benefit of this component-driven approach is testability: because the components can be deployed separately, they can be tested separately.

15.5. Creation of a generic SPARQL endpoint for OAI-PMH repositories

We searched for software that might present an OAI-PMH repository as a SPARQL endpoint. The closest we found was OAI-PMH RDFizer from the Simile project (<http://simile.mit.edu/wiki/RDFizers>). Shortcomings of RDFizer for our purposes are that it does not of itself create a queryable interface for accessing the metadata, and it is not clear to us how well it can deal with non-standard OAI-PMH metadata. Instead, we are planning to construct a generic SPARQL endpoint for OAI-PMH repositories ourselves, using our own instance of EPrints as a testbed. This will provide an example of the lightweight software approach that we espouse for data web development:

The Jena system from HP Labs (<http://jena.sourceforge.net/>) provides two complete applications, Joseki and the Jena model loader, which can run against a common database. We plan first to develop a simple, free-standing local harvester that uses OAI-PMH to collect selected metadata from the EPrints repository on a periodic basis, and present that as a simple RDF file in Notation 3 format (<http://www.w3.org/DesignIssues/Notation3>). The Jena model loader, which has built-in support for Notation 3, will then write the data into a Jena RDF database. Joseki can access the same Jena database (concurrently, if necessary) in response to a SPARQL query, thus providing a SPARQL endpoint for the harvested repository content. In this way, with only a small investment in software development, we immediately provide a mechanism to query across the whole metadata content of the repository, a feature that is not provided by OAI-PMH.

Joseki will thus serve as a *local* per-repository query endpoint, in contrast with systems like OAIster and Intute Repository Search that provide centralised search facilities across metadata harvested from many repositories. Such tooling will create a standard access mechanism for repository data, and opens the possibility of deploying secondary services, such as the semantic browse tool mSpace, across repositories.

This proposal, which involves harvesting and rewriting OAI-PMH metadata into a separate local RDF database, should be clearly distinguished from SPARQL endpoints that are more tightly integrated with the underlying data store and involve query re-writing into the native language of that store (e.g. into SQL for D2R Server that can access relational databases). While involving duplication of data, our approach will have the advantage of being entirely generic for use with any OAI-PMH data source. The metadata rewriting and query rewriting methods (Figure 2, Section 6.3.) are both reasonable approaches, and in these early days of using SPARQL, one of the important outcomes of our proposed work will be to enable performance comparisons between them.

16. PROJECT CONCLUSIONS AND RECOMMENDATIONS

- See [DefiningImageAccess/RecommendationNotes](#)

These have already been presented in the Executive Summary (Sections 2.3 and 2.4, q.v.), and will not be repeated here.

17. GLOSSARY OF NAMES, ABBREVIATIONS AND ACRONYMS

See also: [DefiningImageAccess/Glossary](#). This list excludes acronyms for JISC Projects, which can be resolved at <http://www.jisc.ac.uk/search.aspx>.

AKT	Advanced Knowledge Technologies, http://www.aktors.org/akt/ .
Atom	The Atom syndication format is an XML language used for web feeds, http://en.wikipedia.org/wiki/Atom_(standard) (See also RSS.)
CCLRC	Central Laboratory of the Research Councils, http://www.cclrc.ac.uk/ . CCLRC has recently merged with the PPARC to become the Science and Technology Facilities Council (STFC; http://www.stfc.ac.uk).
CIDOC CRM	The Conceptual Reference Model (ISO International Standard ISO-21127; http://cidoc.ics.forth.gr/working_editions_cidoc.html) developed by the Committee on Documentation of the International Council of Museums (ICOM-CIDOC).
CORDRA Project	Content Object Repository Discovery and Registration/Resolution Architecture, http://www.cordra.net/ .
CrossRef	CrossRef is the official DOI link registration agency for scholarly and professional publications, and operates a cross-publisher citation linking system, http://www.crossref.org/ .
CSS	Cascading Style Sheets, http://www.w3.org/Style/CSS/ .
D2R Server	Database to RDF Server, http://sites.wiwiss.fu-berlin.de/suhl/bizer/d2r-server/ .
DC	Dublin Core, http://dublincore.org/ .
DCC	Digital Curation Centre, http://www.dcc.ac.uk/ .
DOI	Digital Object Identifier, a form of unique identifier widely used in publishing, http://www.doi.org/ .
FRBR	Functional Requirements for Bibliographic Records, http://www.frbr.org/ .
GRDDL	Gleaning Resource Descriptions from Dialects of Languages, http://www.w3.org/2004/01/rdxh/spec .
HTML	Hypertext Markup Language, http://www.w3.org/TR/html4/ .
HTTP	Hypertext Transfer Protocol Overview, http://www.w3.org/Protocols/ ; http://www.ietf.org/rfc/rfc2616.txt .
IBRG	Image Bioinformatics Research Group, http://imageweb.zoo.ox.ac.uk/ .
ILRT	Institute for Learning and Research Technologies at the University of Bristol, http://www.ilrt.bris.ac.uk/ .
INDECS	Interoperability of Data in e-Commerce Systems, http://www.indecs.org/ .
Ingenta	An aggregator for scholarly bibliographic information, now part of Publishing Technology plc, http://www.ingenta.com/ , http://www.ingentaconnect.com/
JISC	Joint Information Systems Committee, http://www.jisc.ac.uk/ .
JPEG	Joint Photographic Experts Group, http://www.jpeg.org/ .

METS	Metadata Encoding and Transmission Standard, http://www.loc.gov/standards/mets/ .
MIBBI	Minimal Information about Biological and Biomedical Investigations, http://mibbi.sourceforge.net/ .
MP3	MPEG-1 Audio Layer 3, http://www.mpeg.org/ .
MPEG-21 DID	MPEG-21 Digital Item Declaration, http://www.chiariglione.org/mpeg/technologies/mp21-did/index.htm .
N3	Notation 3 is a readable alternative to the RDF/XML format for serialized RDF, http://www.w3.org/DesignIssues/Notation3 .
OAI-ORE	Open Archives Initiative Protocol - Object Exchange and Reuse, http://www.openarchives.org/ore/ .
OAI-PMH	Open Archives Initiative - Protocol for Metadata Harvesting, http://www.openarchives.org/OAI/openarchivesprotocol.html .
OAster	OAI-PMH Harvester, http://www.oaister.org/ .
ODL	Oxford Digital Library, http://www.odl.ox.ac.uk/ .
OGSA-DAI	Open Grid Services Architecture – Database Access and Integration, http://www.ogsadai.org.uk .
ORA	Oxford Research Archive, http://ora.ouls.ox.ac.uk/access/ .
OWL	Web ontology language, http://www.w3.org/TR/owl-features/ .
PDF	Portable Document Format, http://www.adobe.com/devnet/pdf/pdf_reference.html , http://en.wikipedia.org/wiki/Portable_Document_Format .
PNG	Portable Network Graphics, http://www.libpng.org/pub/png/ .
PREMIS	PREservation Metadata: Implementation Strategies, http://www.oclc.org/research/projects/pmwg/ .
R4L	Repository for the Laboratory, http://r4l.eprints.org/ .
RDF	Resource Description Framework, http://www.w3.org/RDF/ .
RDFS	RDF Schema, http://www.w3.org/TR/rdf-schema/ .
REST	Representational State Transfer, http://www.ics.uci.edu/~fielding/pubs/dissertation/abstract.htm .
ROAR	Registry of Open Access Repositories, http://roar.eprints.org .
RPC	Remote Procedure Call, http://en.wikipedia.org/wiki/Remote_procedure_call .
RTFC	The Research and Technology Facilities Council (http://www.rtfc.ac.uk), formerly CCLRC.
RSS	Really Simple Syndication (AKA RDF Site Summary, or Rich Site Summary) , http://en.wikipedia.org/wiki/RSS_(file_format) , http://cyber.law.harvard.edu/rss/rss . (See also Atom.)
SCARP	A JISC-funded project of the Digital Curation Centre to determine requirements for preservation of digital data, http://www.dcc.ac.uk/scarp/ .
SCORM	Sharable Content Object Reference Model, http://www.adlnet.gov/scorm/index.aspx .

SKOS	Simple Knowledge Organisation System, http://www.w3.org/TR/swbp-skos-core-guide/ .
SOA	Service Oriented Architecture, http://en.wikipedia.org/wiki/Service Oriented Architecture .
SOAP	Simple Object Access Protocol, http://www.w3.org/TR/soap/ .
SPARC Europe	An alliance of European research libraries, http://www.sparceurope.org/ .
SPARQL	SPARQL Protocol and RDF Query Language, http://www.w3.org/TR/rdf-sparql-query/ .
STFC	Science and Technology Facilities Council, http://www.scitech.ac.uk/ .
SWAP	A Scholarly Works Application Profile, http://www.ukoln.ac.uk/repositories/digirep/index/Eprints Application Profile#Scholarly Works Application Profile
TASI	Technical Advisory Service for Images, http://www.tasi.ac.uk/ .
TRiX	Triples in XML, http://sw.nokia.com/trix/TriX.html .
UDDI	Universal Description, Discovery and Integration, http://www.uddi.org/
UKOLN	UK Office for Library Networking, http://www.ukoln.ac.uk/ .
URI	Uniform Resource Identifier, http://tools.ietf.org/html/rfc3986 , http://www.w3.org/Addressing/ .
URL	Uniform resource locator http://tools.ietf.org/html/rfc1630 , http://www.w3.org/Addressing/ .
VRE	Virtual Research Environment, http://www.jisc.ac.uk/whatwedo/programmes/programme_vre.aspx .
W3C	World Wide Web Consortium, http://www.w3.org/ .
WSDL	Web Service Definition Language, http://www.w3.org/TR/wsdl .
XHTML	Extensible Hypertext Markup Language, http://www.w3.org/TR/xhtml1/ .
XML	Extensible Markup Language, http://www.w3.org/XML/ .

REFERENCES

(The following list has been copied to [DefiningImageAccess/References](#), where it will continue to be added to as our work continues, after the Final Report has been delivered).

- Allinson, Johnston and Powell (2007). A Dublin Core application profile for scholarly works. *Ariadne* **50**, <http://www.ariadne.ac.uk/issue50/allinson-et-al/>.
- Berners-Lee (1989). Information management: a proposal. CERN Report, <http://www.w3.org/History/1989/proposal.html>.
- Berners-Lee (1999). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*, Harper Collins.
- Berners-Lee (2007). Tim Berners-Lee talks about the Semantic Web as a data web (video), Technologyreview.com, <http://www.technologyreview.com/video/semantic>.
- Brickley (2003). Missing isn't broken: data validation and freedom on the Semantic Web. RDFWeb and Friend of a Friend (FOAF) Project Log, <http://rdfweb.org/mt/foaflog/archives/000047.html>.
- Brickley (2005). CheckRDFSyntax and Schemarama Revisited. Danbri's foaf stories: the web, the world, us, you and them, Blog, <http://danbri.org/words/2005/07/30/114>.
- Carroll, Bizer, Hayes *et al.* (2004). Named Graphs, Provenance and Trust. HP Laboratories, Bristol, HPL-2004-57, <http://www2005.org/cdrom/docs/p613.pdf>.
- Connolly (2007). Microformats: what are they, and why should we use them? XML Summer School, Oxford, iCalendar Data W3C Interest Group Note, <http://www.w3.org/2007/07dc-lhr/mf-xmlsum.pdf>.
- Doan and Halevy (2005). Semantic integration research in the database community: a brief survey. *AI Magazine, Special Issue on Semantic Integration, Spring 2005*, <http://pages.cs.wisc.edu/~anhai/projects/schema-matching.html>.
- Emmott (2006). Towards 2020 Science. *Microsoft Research Report*, <http://research.microsoft.com/towards2020science/>.
- Fielding (2000). Architectural styles and the design of network-based software architectures. Chapter 5: Representational state transfer (REST). Ph. D. thesis. Department of Information and Computer Science, University of California, Irvine, <http://www.ics.uci.edu/~fielding/pubs/dissertation/abstract.htm>.
- Franklin, Halevy and Maier (2005). From databases to dataspace: a new abstraction for information management. *ACM SIGMOD Record*.
- Grossman (2003). Integrating distributed bioinformatics data using data webs. *O'Reilly Bioinformatics Technology Conference: Chicago, Illinois, Add.*
- Halevy, Franklin and Maier (2006). Principles of dataspace systems. PODS '06, Chicago, Illinois, USA, ACM 1595933182/06/0003.
- He, Patel, Zhang *et al.* (2007). Accessing the deep web. *Communications of the ACM* **50** (5): 94-101, <http://portal.acm.org/citation.cfm?doid=1230819.1241670>.
- Hughes, Mills, de Roure *et al.* (2004). The semantic smart laboratory: a system for supporting the chemical eScientist. *Organic and Biomolecular Chemistry* **2**: 1-10.
- Klyne (2006). Inter-portlet communication considered harmful. OSS Watch Wiki, <http://wiki.oss-watch.ac.uk/InterPortletCommunicationConsideredHarmful>.
- Kondylakis, Doerr and Plexousakis (2006). Mapping language for information integration. Technical Report 385, ICS-FORTH, http://www.ics.forth.gr/isl/publications/paperlink/Mapping_TR385_December06.pdf.
- Lagoze and Van de Sompel (2007). Compound Information Objects: The OAI-ORE Perspective. Open Archives Initiative – Object Reuse and Exchange, <http://www.openarchives.org/ore/documents/CompoundObjects-200705.html>.
- 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.

- Matthews, Portwin, Pepler *et al.* (2007). Cross-linking and referencing data and publications in Claddier. UK e-Science All Hands Meeting September 2007, Nottingham, <http://epubs.cclrc.ac.uk/work-details?w=37696>.
- Mika (2005). Ontologies are us: a unified model of social networks and semantics. International Semantic Web Conference ISWC2005.
- Muggleton, Vinge, Szalay *et al.* (2006). 2020 Computing. *Nature* **440**: 409-419.
- Neumann and Quan (2006). A Semantic Web dashboard for drug development. *Pacific Symposium on Biocomputing* **11**: 176-187.
- 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>.
- Pryor (2006). Attitudes and aspirations in a diverse world: the Project StORe perspective on scientific repositories. 2nd International Digital Curation Conference, Glasgow, <http://jiscstore.jot.com/WikiHome/DisseminationPages/IDCC2%20-%20Paper.doc>.
- Shotton, Rodriguez, Guil *et al.* (2002). A metadata classification schema for semantic content analysis of videos. *Journal of Microscopy* **205**: 33 -42.
- Swan and Awre (2006). Linking UK Repositories: Technical and organizational models to support user-oriented services across institutional and other digital repositories. JISC-commissioned scoping study report, http://www.jisc.ac.uk/uploaded_documents/Linking_UK_repositories_report.pdf.
- Van de Sompel, Bekaert, Liu *et al.* (2005). aDORe: a modular, standards-based digital object repository. *The Computer Journal Advance Access*, <http://comjnl.oxfordjournals.org/cgi/rapidpdf/bxh114v1.pdf>.
- Zhao (2007). Semantic matching across heterogeneous data sources. *CACM* **50** (1): 45-50, <http://portal.acm.org/citation.cfm?id=1188913.1188916>.

18. INTRODUCTION TO APPENDIX A

The following report, constituting Appendix A of this Final Report, was written by Julie Allinson, the UKOLN Consultant Partner on the JISC *Defining Image Access* Project, who participated in and contributed to all four of the formal project meetings (http://imageweb.zoo.ox.ac.uk/wiki/index.php/Defining_Image_Access#Project_meetings), and whose consultancy agreement included contributing to the final report from the UKOLN perspective.

Her report was written in response to the following invitation on 23 May 2007 from David Shotton, the *Defining Image Access* Project's P.I:

“Since you already have such an intimate knowledge of JISC repository activities, you are ideally placed to write sections of the *Defining Image Access* Project Final Report that deal (a) with the relationship between this project and other activities of the JISC Repositories and Preservation programme, and (b) with the relationship between any image webs that might subsequently be developed for university repositories and other JISC services such as the Intute Repository Search.

Specifically, I would like your report to include four sections:

- (a) A description of the relationship between the *Defining Image Access* Project and other existing JISC activities.
- (b) A description of the relationship between the *Defining Image Access* Project and any other relevant activities of which you are aware that involve images and/or are being undertaken by or on behalf of university repositories.
- (c) The potential for future data webs to provide interoperability between images and research publications in institutional repositories, and the relationship between such data webs and the Intute Repository Search service.
- (d) Specification of what we would need to do to ensure that our future activities, specifically in developing such future data webs:
 - (i) are compliant with the JISC e-Framework and Information Environment Architecture, and the JISC Information Environment Technical Standards (<http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/standards/>), or, conversely, the degree to which those standards fail to encompass the Semantic Web approach that we are proposing;
 - (ii) can be properly registered with the JISC Information Environment Service Registry (IESR, <http://www.iesr.ac.uk/>), and the IE Metadata Schema Registry (IEMSR, <http://www.ukoln.ac.uk/projects/iesmr/>); and
 - (iii) might contribute to the development of the proposed JISC Application Profile for Images, that TASI has been commissioned to develop, based upon the excellent JISC Application Profile for Scholarly Works that you and Andy Powell authored.

What will be so valuable is your objective and independent assessment of our own ideas.”

Julie's report was written independently from this *Defining Image Access* Final Report, and was submitted on 25 June 2007. Although we initially intended to combine its content into the body of the Final Report, we subsequently decided to maintain it as a separate Appendix, despite the fact that this results in some duplication when discussing third-party projects. This is because it provides a valuable ‘expert outsider’s’ perspective on the project, which would have lost much of its clarity and impact if it had been merged with the body text.

19. APPENDIX A: AN INDEPENDENT COMMENTARY BY JULIE ALLINSON

THE DEFINING IMAGE ACCESS PROJECT AND ITS RELATIONSHIP WITH OTHER JISC ACTIVITIES

by Julie Allinson, Repositories Research Officer
UKOLN, University of Bath, Bath BA2 7AY, United Kingdom
Tel: +44 (0) 114 2486457, +44 (0) 1225 386580
j.allinson@ukoln.ac.uk; <http://www.ukoln.ac.uk/ukoln/staff/j.allinson/>

Table of Content

1. Images in an institutional repositories context	50
2. The data webs concept	51
3. Joining up with JISC services, projects and activities	51
4. Discovery services	51
5. Linking research data and publications	52
6. Repositories-related projects	53
7. JISC Services	55
8. Beyond JISC, connections with the wider world	56
9. Data webs and interoperability	56
10. Metadata and object modelling issues	57
11. Conclusion	58

1. Images in an institutional repositories context

A simple search of the JISC web site for ‘images’ returns 20 pages, for ‘image’ 353. Suffice to say that images play a significant part across many of the activities and funding of the JISC (Joint Information Systems Committee - <http://www.jisc.ac.uk/>). In research, images are truly cross-discipline. They illustrate the research publications created by researchers; they are collected as research data, stored in databases and file stores, and analysed, compared and examined; they are displayed at exhibitions, included in presentations, used in teaching, and ‘snapped’ to create a record of events. Yet one might ask, what is an image? Restricting ourselves to digital representations, we are still faced with a range of image formats (jpeg, gif, png etc.), types (computer-generated graphics, photographs, digitised images of physical objects, for example) and purposes (art, scientific study, geographic location, mapping, news, comparative analysis etc.).

With images the importance of context and metadata cannot be underestimated. Images are not self-describing, they come with no ‘full-text’ - content providers and aggregators depend on descriptive metadata to add semantic richness. Technical metadata can be embedded within images themselves, allowing richer machine to machine interactions. Descriptive information must be created by people through formal cataloguing or informal tagging and social annotation techniques, or by automated means.

The JISC *Defining Image Access* requirements analysis project (DIA) has shown that this metadata includes both generic metadata common to most images and domains, and discipline-specific metadata. The former offers a common access vocabulary, the latter offers necessary specificity, but if used wrongly, can cause confusion and reduces interoperability. Where images carry meaning specific only to the creator of that image, sharing, re-use and re-interpretation by individuals, communities and disciplines adds additional meaning, which can only become accessible if this information, too, is captured and shared.

The Defining Image Access project has considered this range of issues carefully. It is born out of the work of the BioImageWeb Consortium, an informal group of academic researchers and publishers exploring options for enhancing the value of biological research images, in journal publications or elsewhere online, by making them interoperable using a data web approach. The consortium brings considerable and varying expertise to the project. Although it has its roots in images related to biology and zoology, the DIA project focus is truly cross-disciplinary and planted firmly at the door of institutional repositories. This is a crucial aspect of the project for several reasons:

1. It highlights the potential dislocation between institutional ‘eprint’ repositories and other digital resource collections that may exist within universities and elsewhere. At a human level, identifying different collections and repositories can be difficult; for machines, this can be similarly problematic. This is part of the broader interoperability domain in which the JISC Information Environment is working. Integrating systems, sharing content in a standardised way and, through these, increasing the efficacy of machine-machine and human-machine interactions are key goals of the Information Environment.
2. It demonstrates that, in many cases, repository systems have been built to support a fairly specific type of resource and are only now beginning to fully explore the complexity of the resources they house. Images, in many cases, are embedded within pdf or other document types, rendering them impossible to share in the current information landscape. DIA has found that institutional repositories contains relatively few image collections – a surprising finding which is significant for repository development.
3. Establishing the links between one resource and other resources can be easy using simple web linking techniques, but for images, where links cannot be easily embedded within the data itself, establishing links is a significant issue. Understanding the nature of those links

too, is a particular problem across the Web and is currently being explored by the OAI-ORE project, covered later in this section.

4. Finally, it raises awareness that repositories, as relatively new data stores and services, must be clear on where they fit within the institutional, subject and national/international context and of how opportunities for leveraging new methods of sharing and accessing content with existing and emerging services might be exploited.

DIA, with its 'data webs' concept, is exploring key elements in the information landscape, outlining methods for offering 'join-up' and increased access to content using lightweight approaches that are already in use on the Web.

2. The data webs concept

Is the idea of a 'data web' just another search service? Don't we have enough of those already? By undergoing a rigorous requirements gathering process, DIA has sought to answer these questions. Data webs are based on the principal of distributed web publication of data and accompanying metadata, together with lightweight harvesting and aggregation of the metadata to a form that can be browsed and searched from a single point of entry, with direct links back to the original data sources to allow delivery. In the context of the Information Environment and in light of reports such as the Repositories Review and CLIC, it seems that there is a genuine use case to support the creation of data webs, both in offering enriched services to the end user and in demonstrating the viability of combining Web 2.0, Semantic Web and institutional approaches. Key aspects of the concept offer potential for added-value and new services; these include: 1) enriching metadata with annotations, 2) providing machine access to data webs for other services, 3) offering a co-referencing service and 4) mapping different metadata schemas into a single searchable RDF graph. By leaving control of data and metadata with the data provider, and by facilitating discovery through linking back to the source repository, data webs remain lightweight and dynamic and need not be burdened by the complexities of access control and data management. They don't lock data up or duplicate it; rather they enhance access and potentially increase usage, improving business models and sustainability. Social and policy issues regarding trust and data ownership, although important, are therefore not the responsibility of data webs.

The feasibility, success or future sustainability of data webs is not guaranteed. DIA acknowledges that the proposed services and mapping activities may demand significant effort and may only provide an interim patch until practices have improved. Projects such as this, though, have a significant role to play in improving practices. As a starting point, the DIA project has established that the data webs idea has potential value, and has suggested a technical architecture. Implementation and testing of this, through developing data webs and supporting tools is the obvious next step.

3. Joining up with JISC services, projects and activities

From its outset, DIA has been committed to drawing on expertise and connecting with other projects working across a similar space. Its partners can all demonstrate considerable awareness of and involvement in other JISC projects. DIA is cognisant of the broader JISC vision for an information environment and its goals are in line with those of the e-framework, as discussed in a later section. What follows is a brief review of JISC projects, JISC services and other JISC work that has informed the project throughout its life and may be further explored in any continuing work.

4. Discovery services

A number of projects funded within the JISC **FAIR**, **Digital Repositories** and **Repositories and Preservation** programmes are developing services to aggregate and search resources. PerX (<http://www.engineering.ac.uk/>) and IRIScotland (<http://cdlr.strath.ac.uk/irisotland/>) both offer important experiences and lessons learnt, in addition to their demonstrator services. PerX

(<http://www.icbl.hw.ac.uk/perx/>) offers an interesting dimension of subject-specificity and has highlighted significant issues with metadata quality and in accessing subject resources and the use of subject classifications and ontologies. Both are focussed primarily on textual research publications, rather than images.

Of particular importance is the newly-funded **Intute Repository Search** project (<http://www.intute.ac.uk/projects.html>). A relationship with this project has been established by DIA and discussions explored how a 'data web' might impact on the future development of this project are ongoing. Like PerX and IRIScotland, the initial focus of Intute's search project is on research texts, or eprints, as their early pilot service demonstrates - <http://irs.ukoln.ac.uk/>. Planned future work will look at added-value services and extending the coverage to different resource types. There is a clear synergy with DIA - joining up approaches and further discussions are clearly going to be of great value. Development of any supporting services, such as the co-reference or metadata mapping services proposed by DIA would be mutually supportive and should be further explored.

Intute (<http://www.intute.ac.uk>), the parent service of the search project also offers possible synergy through its development of semantic personalisation tools. Additionally, its collection of resource descriptions offer a wealth of links to image resources.

5. Linking research data and publications

The boundary between images and research data is somewhat blurred. Images embedded in research publications could be viewed as purely illustrative, with no further intrinsic value. Yet if viewed as research data, this is clearly not the case, as the images themselves become 'data' with the potential for re-use and sharing. There has been considerable work in the area of 'research data' in recent years, of which DIA is aware. Images as data feature in the significant consultancy report by Dr Liz Lyon 'Dealing with data' (http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf). Lyon analyses the activities of a number of existing data repositories and notes, in particular, that both the Arts and Humanities Data Service and UK Data Archive manage image resources as data. Lyon also refers to the DCC **SCARP** (<http://www.dcc.ac.uk/scarp/>) project which is engaging with "particular disciplines to learn more about data curation practice 'at the coalface'" (page 48), looking at curation in its broadest sense – data deposit, management, sharing, re-use and preservation. Within SCARP, the **Image Store project** (http://www.dcc.ac.uk/scarp/#image_store), looking at requirements for effective curation of scientific research images, video and associated data from the biological domain, has a strong connection established with DIA, since both are managed by the same PI.

DIA is particularly interested in issues relating to the linking of image data with publication, and a number of JISC projects are developing experimental services for searching across research publications and research data, of relevance to DIA:

- The **CLADDIER** project (<http://claddier.badc.ac.uk/trac>) have considered issues surrounding the citation of data and have developed an experimental search service to link data with publications and the data they cite – <http://isegserv.itd.rl.ac.uk/claddier/search/single/> In addition, they are also working on a 'ping' tool to offer a kind of trackback service between cited and citation. This ties in closely with the DIA proposal for a co-reference service. Members of the CLADDIER are DIA consultant partners and have been involved in DIA project meetings.
- The **eBank** project (<http://www.ukoln.ac.uk/projects/ebank-uk/>) too has developed an experimental search service in the context of crystallography data (including images) - <http://ebank.ukoln.ac.uk/>
- The **StORe** project (<http://jiscstore.jot.com/WikiHome>), studying researcher attitudes and behaviour in this area, will also demonstrate bi-directional links between the UK Data

Archive (source repository) and Research Articles Online, an output repository at the London School of Economics).

6. Repositories-related projects

DIA are tapping into a variety of cross-cutting issues affecting repositories. Use of lightweight standards and semantic web tools, availability of content, metadata creation, interoperability and handling complex objects. Connections with other JISC projects have been made throughout the DIA project. Various new projects within the Repositories and Preservation Programme offer additional future possibilities. These include:

Existing projects:

- The **ASK** project (<http://ask.oucs.ox.ac.uk/index.html>) based at Oxford has taken a lightweight development approach to building a repository system, drawing on existing standards and taking a service-oriented approach. Standards implemented include SRU/W, OpenURL and Creative Commons licensing. DIA has already made contact with this project to share experience.
- **CLIC** - <http://www.oucs.ox.ac.uk/ltg/projects/clic/>. Now complete, the CLIC study, also based at Oxford, reviewed the growth of community-owned digital image collections, surveyed barriers to building image collections, and made recommendations for national initiatives to help sharing and embedding collections in the FE/HE sectors. This report is directly relevant to the DIA project and ongoing contact with its authors has proved useful. The data webs concepts offers a mechanism for realising some of the CLIC recommendations. CLIC offers a source of information about existing image collections.
- **Dictate** (http://www.jisc.ac.uk/whatwedo/programmes/programme_pals2/synthesis/projects/dictate.aspx). This project offers a mechanism for integrating institutional repositories with social bookmarking through the development of the Connotea tagging tool for use in the Eprints software. This tools allows users to tag, see tags assigned by others and view items related to a deposited eprint. It offers potential for gathering additional metadata about images from new sources and options for enriching discovery. Other activities, such as the Rich Tags and the EnTag projects offer interesting, more formalised possibilities for tagging and classifying resources. All of these projects are of particular interest to DIA with regards their different and new approaches to metadata capture.
- **MIDESS** (<http://www.leeds.ac.uk/library/midess/>) has looked at institutionally managed image collections and has produced a number of reports of interest to DIA, particularly those relating to image metadata.
- **Terminology Registry Scoping Study** (<http://www.ukoln.ac.uk/terminology/JISC-review2006.html>). This report reviews vocabularies of different types, best practice guidelines, research on terminology services and related projects. It discusses possibilities for terminology services within the JISC Information Environment and eFramework, of relevance to DIA.

New projects:

- The **Art** (http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres/tools/art.aspx) project is of future interest as it plans to create an ontology-based tool for the semantic annotation of papers stored in digital repositories. The ART tool will annotate not only the data and metadata for a paper, but also elements of the scientific investigation itself, and will also be able to aid in the expression of research results in a semantic format. Although this project has no specific focus on image data, its use of semantic methods is significant.

- **BID** (http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres/repositories_sue/bid.aspx) – Bridging the Interoperability Divide – is a new project based at Oxford that plans to demonstrate repository interoperability. Included within its remit are harvest and search providing opportunities for synergy with the data webs approach, and the DIA and BID team are in continuing dialogue.
- **CAIRO** (<http://cairo.paradigm.ac.uk/about/index.html>) - This project, also based in Oxford, is addressing a variety of curation issues posed for repositories during the ingest process. Its emphasis is on complex inter-related collections of objects and preservation, as well as access, metadata requirements. Although currently its activities are beyond those being analysed by DIA, tools developed by this project might be significant in the future. Images are frequently part of collections, with specific provenance and hierarchies. Improving metadata collection techniques at the point of ingest is of significant interest.
- The **DexT** (http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres/dext.aspx) project is developing data exchange tools for use with survey data and qualitative research data. These tools, based on XML/RDF schema, may be of future use in the context of data webs.
- Enhanced Tagging for Discovery (**enTag**) will build a semantic interoperability demonstrator combining controlled and folksonomy approaches to support resource discovery in repositories and digital collections. Working with authors at the point of deposit (CCLRC) and researchers at the point of discovery (Intute). DIA has identified the need for tools that support the metadata creation and ingest stages, to reduce the need for heavyweight post-processing and enhancement at the point of access and search. Parallels with DIA are clear and the usefulness of such techniques within subject-specific data webs is of particular interest. This project has not yet commenced. Contact has already been established with UKOLN, leaders of this project.
- **Rich tags** - <http://www.mspace.fm/projects/richtags/>. Like enTag, this project is looking at the combination of informal tagging processes with formal metadata collection techniques. Tools from this project will offer significant possibilities for metadata capture in any data web development. Contact has already been established with this project.
- **RIDIR** project (<http://www.hull.ac.uk/ridir/>). This is a new project investigating the benefits and requirements for clear use of persistent identifiers for interoperability. There is a potential link here with the data webs concept, particularly its outline for a co-reference service.
- The **SAFIR** project (Sound Archives Film Images Repository; http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres/repositories_sue/safir.aspx) will create a repository of image, audio and video materials. A future connection with this project could yield a useful source of content.
- **SOURCE** (<http://www.source.bbk.ac.uk/>) - Sharing Objects Under Repository Control with Everyone - is building bulk-migration tools for use between different repositories using the OKI OSIDs. Although the focus of data webs is on search, rather than object transfer, the management of image collections and the tools being developed by this project are of interest.
- **SWORD** (<http://www.ukoln.ac.uk/repositories/digirep/index/SWORD>) is a short project developing a lightweight specification for repository deposit. A primary use case of this project is to make deposit to multiple repositories a reality. Its concerns are more technical than metadata-related, yet its use of the Atom and the Atom Publishing Protocol are of significant interest to DIA, particularly in the realm of ingesting materials, but also for syndication of content.

7. JISC Services

Sources of content

There are various JISC services (www.jisc.ac.uk/whatwedo/services) which might act as sources of image content in a future data web project.

- The **Arts and Humanities Data Service** (<http://ahds.ac.uk/>) hosts a wide range of image collections, whose use is at present focussed on researchers within arts and humanities disciplines. Leveraging this content within data webs and potentially broadening its use to other disciplinary audiences is an interesting use case. Licensing issues would need exploring. The AHDS is also leading the meta tools project (<http://ahds.ac.uk/about/projects/metatools/index.htm>) to investigate methods of generating metadata. This is also of significant interest to the work of DIA and data webs. [This service is scheduled for closure in March 2008.]
- The **UK Data Archive** (<http://www.ukda.ac.uk>) contains some multimedia collections. Again, this offers a source of content for data webs.
- The **Education Image Gallery** (<http://edina.ac.uk/eig/>) is a licensed image service run by Edina. Whilst its contents are not freely available it is another example of a significant image collection made available through funding from JISC.
- **Jorum** (<http://www.jorum.ac.uk>), the national learning object service, contains some image resources, many of which are likely to be embedded within content packages or documents. Although its focus is on teaching, rather than research, Jorum might offer a source of image materials for use within data webs. Aggregating images and licensing issues would again need consideration.
- **Edina** (<http://www.edina.ac.uk/>), run various projects and services, mainly in the realm of geospatial data, which might offer image content or useful tools, including:
 - The Visual & Sound Materials portal scoping study & demonstrator project (<http://www.edina.ac.uk/projects/vsmportal/>)
 - Go-Geo! (<http://www.gogeo.ac.uk/>)
 - GRADE project (<http://edina.ac.uk/projects/grade/>)
 - geoXwalk (<http://www.geoxwalk.ac.uk/>)

Machine services

The **Shared Services Review** (<http://www.jisc.ac.uk/media/documents/programmes/capital/jisc-sis-report-final.pdf>) looks at existing components of the JISC Information Environment, offering a rich source of information about machine services which might be used in the development of data webs. Of particular interest are IESR (<http://iesr.ac.uk/>) and the IEMSR (<http://www.ukoln.ac.uk/projects/iemsr/>), both are considered later in this report.

Advisory services

- TASI (<http://www.tasi.ac.uk/>)
- OSS Watch (<http://www.oss-watch.ac.uk/>)
- UKOLN (<http://www.ukoln.ac.uk>)
- AHDS (<http://www.ahds.ac.uk>)
- JISC CETIS (<http://www.cetis.ac.uk/>)

8. Beyond JISC, connections with the wider world

A wealth of image collections - scholarly, historical, archival, commercial and more - exist and can be discovered via the Web. For understandable IPR, commercial and confidentiality reasons, many are subject to access restrictions. Yet, for many, lightweight approaches to both the technology and policy issues can turn what might otherwise become closed and underused 'silos' of content into reusable scholarly resources.

One significant example in the social realm, worthy of further consideration is **Flickr** (<http://www.flickr.com/>). Flickr is an international photo sharing service which allows programmatic access to its API for non-commercial use, allowing developers to build a wealth of tools to leverage and enrich the content in the Flickr database. Flickr also demonstrates many of the features that data webs might wish to support or explore – user tagging, annotations, geographic locations, licensing, access restriction, collection building and image sizing. In the realm of scholarship, the images and their use may be of a different nature, placing additional demands on metadata and interoperability, yet Flickr's principles of lightweight technology and openness offer much that might be drawn upon in the research domain. Its limitations with regards metadata and the inconsistent use of tagging are issues and mirror, to a degree, those faced in education.

Picture Australia, a search service for Australian historical picture collections, has used Flickr to build a community image collection to demonstrate how formal collections can co-exist with and be enriched by community-led image sharing (<http://www.pictureaustralia.org/Flickr.html>). It provides an excellent model for an image data web of which the DIA team are aware.

Beyond the UK, where Intute, PerX and IRIScotland are developing search services, there exist further examples, such as BASE (<http://www.base-search.net/>) and OAIster (<http://quod.lib.umich.edu/>). The latter allows searches of 'image' as a resource type, but cannot offer richer functionality that might be facilitated by an image and community-specific data web.

9. Data webs and interoperability

The JISC Information Environment

Data webs sit across the fusion and presentation layers envisaged by the **JISC Information Environment** (<http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/>), and are in line with the technical architecture outlined. The IE technical standards, required for services being built within the context of the IE, can be used to guide the standards chosen by any data web development, although it is not the intention of the IE to dissuade leading edge projects from testing new standards. Standards of relevance include SRU/W, Z39.50 and OAI-PMH for search Dublin Core, ATOM, RDF/RDFS, OWL, SKOS for metadata, and HTTP POST, Atom Publishing Protocol and WebDav for deposit. The **JISC Standards Catalogue** (<http://standards.jisc.ac.uk/catalogue/Home.phtml>) offers further guidance, as does Andy Powell's briefing paper "A 'service oriented' view of the JISC Information Environment" (<http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/soa/>).

SPARQL (<http://www.w3.org/TR/rdf-sparql-query/>), not yet covered in either source of guidance, although implemented by the IEMSR, is one example where DIA can directly input into the uptake and testing of new standards. DIA is strongly supporting the implementation of SPARQL and the desirability of SPARQL end points on institutional repositories. The broad vision of the JISC IE (<http://www.jisc.ac.uk/ie/>) "to allow discovery, access and use of resources for research and learning irrespective of their location" is well reflected in the DIA vision, whereby resources can be stored and managed heterogeneously and accessed by users via subject-specific data webs.

Any data webs that are built as a result of the DIA requirements exercise should be registered with the **Information Environment Service Registry** (<http://iesr.ac.uk/>) using the mechanism identified by the IESR web site. IESR offers a machine-to-machine tool for promoting use of resources developed in within the IE.

The JISC/DEST E-Framework for Research and Education

The **e-Framework for Education and Research** (<http://www.e-framework.org/>) is an initiative by the JISC and Australia's Department of Education, Science and Training (DEST). Its goal is to facilitate technical interoperability in education and research through service-oriented approaches. It defines the processes, or service genres, required to perform a particular abstract function (e.g. search) and the service expressions that can be implemented (e.g. Z39.50). There is no mandatory requirement attached to using the e-Framework, yet JISC encourages all development activity to consider a service-oriented-approach and to use the standards ratified by the e-Framework. Such an approach is cognisant with the DIA project technical architecture where the onus focuses on using existing standards and tools and on building lightweight web services and tools rather than creating monolithic applications.

The data-web concept could be a useful addition to the e-Framework and could be developed into a Service Usage Model (SUM). This would entail some analysis into identifying and then either writing or re-using Service Genres or Service Expressions. This kind of analysis would be useful for the e-Framework in providing a SUM which utilises the lightweight service-oriented approaches at the core of the e-Framework, a SUM which would be re-usable by others. It would also help to identify gaps in documented service genres and expressions and potentially would provide documentation for additional services. For data webs, inclusion in the e-Framework in this way would give them access to resources, expertise and would provide a foundation for discovering services and documenting the data web model.

10. Metadata and object modelling issues

Descriptive image metadata, metadata mapping, enriching metadata, metadata quality – these are all issues that the DIA study has been considering.

DIA is also aware of planned JISC work to develop a Dublin Core application profile for images, following similar work on an application profile for scholarly works (<http://www.ukoln.ac.uk/repositories/digirep/index/SWAP>), based on Dublin Core (<http://www.dublincore.org/>) and, in particular, on the richer functionality supported by the Dublin Core Abstract Model. There already exist a number metadata standards for images. The JISC activity plans to improve interoperability by identifying a profile that is lightweight enough for repositories to expose compliant metadata easily, yet offers richer functionality that services like data webs could make use of. By using Dublin Core, a format widely implemented within the semantic web, the profile will be well-placed to support the goals outlined by DIA and can be expressed in RDF/XML. The DIA project has been exploring its own metadata requirements and could usefully feed these to the application profile developers. DIA should be represented informally on any discussion group related to the images application profile. Any work to specify a minimal metadata set for data webs should be carried out with awareness of the above activity.

Regarding domain-, or discipline-specific metadata, DIA has made clear that there is a requirement for supporting very specific application profiles within domain-specific data webs. A number of mechanisms exist to support this, e.g. exposing richer metadata formats over OAI-PMH, or offering links to additional metadata records, yet the biggest issues are likely to be those of obtaining community agreement rather than technical implementation. Where such standards already exist within communities, experimentation with sharing domain-specific metadata alongside more generic descriptors has been identified as a beneficial aspect to future data web development. Similarly, where data webs collect annotations, storing and sharing these and maintaining links with original data have been flagged as important issues.

The **IE Metadata Schema Registry** (IEMSR, <http://www.ukoln.ac.uk/projects/iemsr/>) is a significant IE resource in the area of metadata schema development. Once the service becomes fully available, any metadata profile created through JISC-funded activity can be developed using the IEMSR tools and should be deposited into the IEMSR to facilitate sharing across the community.

There is potential for join-up between IEMSR and the proposed data web schema alignment service. Shared schemas and ontologies are of paramount importance for data webs, mapping between generic elements and extending functionality using domain-specific metadata. Discussions would be needed to establish how the IEMSR and the schema service outlined by DIA could work together.

The **OAI-PMH** (<http://www.openarchives.org/>) protocol for metadata harvesting is identified as an essential protocol within the JISC IE technical architecture. Widely used by repositories, this standard has made metadata exchange a reality, yet metadata quality issues, caused largely by the use of oai_dc, are increasingly apparent for those wishing to utilise harvested metadata in new ways. DIA has explored such metadata quality issues and is beginning to examine the mechanism for supporting other metadata formats. Recognising that current OAI-PMH limitations have less to do with technology and more to do with community agreement and education is an important step. The DIA team has been at pains to stress that the better the metadata ingested at the point of data creation in the research lifecycle, the more functionality can be offered at the point of dissemination.

Mechanisms for modelling and describing resources exist and have been explored by DIA, for example CIDOC-CRM, the Dublin Core Abstract Model and OAI-ORE. Standards such as METS, MPEG DIDL, RDF and ATOM too offer possibilities for data exchange and sharing. Awareness of the wide range of existing format and standards is essential for interoperability.

From the creators of OAI-PMH, a new project is looking beyond metadata to the exchange and reuse of objects themselves, along with issues of object modelling. **OAI-ORE** (<http://www.openarchives.org/ore/>) is in its relative infancy at present, yet already its potential impact for repositories and service providers is established. The founding principle of the project is to define our understanding of compound information objects and facilitate the sharing of these, and their constituent parts, using the existing web architecture. Fundamental to the project is the commitment to enrich exchanges with information that can outline the boundary of compound objects and identify the relationships between their components, the nature of these relationships (e.g. hasPart, cites etc.) and the context in which an image is used (e.g. as part of a sequence of images, in a research publication). From this, new services will develop to enrich the interoperability layer. There is a clear synergy with the goals of DIA and any future work on data webs may exploit the richness that ORE promises, and the DIA PI is in contact with Herbert Van de Sompel to develop collaboration.

11. Conclusion

As a short requirements analysis activity, the Defining Image Access Project has successfully demonstrated that there is potential for image data webs across education and research. On a technical level the project has highlighted some of the existing lightweight standards that could support such an activity and has gathered hands-on experience of particular repository software in order to explore the management of image collections on the ground. The approach taken by this project has been inclusive and exhaustive, seeking at all times to explore new standards and technologies, particularly through developing links with JISC and other project work in the area. This has been a particularly strong point for the project, as it has already built up a knowledge base and a set of relationships that will strongly support any future work on data webs. Running concurrently with this landscape-wide survey, the project has been defining requirements for data webs and has come up with an architecture informed by the work mentioned above. The vision is forward looking and ambitious, relying wherever possible on standards and services that already exist, whilst outlining gap areas where new service and standards development are required. The project has made some unexpected findings and its outputs are perhaps not what were anticipated at the start. Being responsive and adaptable is another key strength for the project, working through a new concept and dealing with the unexpected requirements that take precedence over the anticipated ones. One potentially weak area for the project is that, as a technical requirements

analysis, it has not looked in any detail at the non-technical issues, such as IPR, policy and culture, which can often prove to be the bigger barriers to development. Issues relating to the availability and quality of data and metadata from source repositories have been highlighted and might be barriers to future work. Such technical and non-technical barriers may, in reality, prevent the vision and architecture proposed from being fully realised. By having such a requirements gathering project these issues have been considered, yet only by developing data webs can they be truly tested.

From the UKOLN perspective, both in terms of its role in the wider information environment and through its links to various projects, most notably the Intute Repository Search project, we view this Defining Image Access Project and the data webs concept favourably, and believe that the project has been useful.

20. APPENDIX B: THE SCHOLARLY WORKS APPLICATION PROFILE FOR THE *DEFINING IMAGE ACCESS PROJECT* FINAL REPORT

by **Julie Allinson**, Repositories Research Officer, UKOLN
University of Bath, Bath BA2 7AY, United Kingdom
Tel: +44 (0) 114 2486457, +44 (0) 1225 386580
j.allinson@ukoln.ac.uk; <http://www.ukoln.ac.uk/ukoln/staff/j.allinson/>

and

David Shotton, The Image Bioinformatics Research Group, Department of Zoology
University of Oxford, South Parks Road, Oxford OX1 3PS, UK
Phone: +44 (0) 1865 271193
david.shotton@zoo.ox.ac.uk

A Scholarly Works Application Profile (SWAP; http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile#Scholarly_Works_Application_Profile), conforming to the standard established by (Allinson *et al.*, 2007), provides metadata describing this *Defining Image Access Project* Final Report in machine-readable XML format.

The XML version of the Application Profile exists as a separate data file accompanying this report (http://imageweb.zoo.ox.ac.uk/pub/2007/DefiningImageAccess/FinalReport/DIA_FR_SWAP_2007-08-17.xml). Additionally, although designed for ease of comprehension by computers rather than humans, it is listed here for completeness.

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE descriptionSet (View Source for full doctype...)>
<epdcx:descriptionSet xsi:schemaLocation="http://purl.org/eprint/epdcx/2006-11-16/
http://purl.org/eprint/epdcx/xsd/2006-11-16/epdcx.xsd" data-
view:transformation="http://purl.org/eprint/epdcx/xslt/2006-11-16/epdcx2rdfxml.xsl"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:xs="http://www.w3.org/2001/XMLSchema"
xmlns:epdcx="http://purl.org/eprint/epdcx/2006-11-16/" xmlns:data-
view="http://www.w3.org/2003/g/data-view#">
<epdcx:description epdcx:resourceURI="INSERT-uri-for-report-as-work">
<epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/type"
epdcx:valueURI="http://purl.org/eprint/entityType/ScholarlyWork" />
<epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/title">
<epdcx:valueString>Images and Repositories: Present Status and Future
Possibilities. JISC Defining Image Access Project Final Report</epdcx:valueString>
</epdcx:statement>
<epdcx:statement epdcx:propertyURI="http://purl.org/dc/terms/abstract">
<epdcx:valueString>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.
The 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; describes the project's achievements; 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.</epdcx:valueString>
</epdcx:statement>
<epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/creator">
```

```

    <epdcx:valueString>Shotton, David</epdcx:valueString>
  </epdcx:statement>
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/creator">
    <epdcx:valueString>Zhao, Jun</epdcx:valueString>
  </epdcx:statement>
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/creator">
    <epdcx:valueString>Klyne, Graham</epdcx:valueString>
  </epdcx:statement>
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/creator">
    <epdcx:valueString>Allinson, Julie</epdcx:valueString>
  </epdcx:statement>
  <epdcx:statement epdcx:propertyURI="http://purl.org/eprint/terms/affiliatedInstitution">
    <epdcx:valueString>University of Oxford</epdcx:valueString>
  </epdcx:statement>
  <epdcx:statement epdcx:propertyURI="http://purl.org/eprint/terms/affiliatedInstitution">
    <epdcx:valueString>University of Bath</epdcx:valueString>
  </epdcx:statement>
  <epdcx:statement epdcx:propertyURI="http://www.loc.gov/loc/terms/relators/FND">
    <epdcx:valueString>Joint Information Systems Committee</epdcx:valueString>
  </epdcx:statement>
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/identifier">
    <epdcx:valueString
      epdcx:sesURI="http://purl.org/dc/terms/URI">http://imageweb.zoo.ox.ac.uk/pub/200
      7/DefiningImageAccess/FinalReport/</epdcx:valueString>
    </epdcx:statement>
  <epdcx:statement epdcx:propertyURI="http://purl.org/eprint/terms/isExpressedAs"
    epdcx:valueRef="dia-final-report-1" />
</epdcx:description>
<epdcx:description epdcx:resourceId="dia-final-report-1">
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/type"
    epdcx:valueURI="http://purl.org/eprint/entityType/Expression" />
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/terms/description">
    <epdcx:valueString>This is the final version of the Defining Image Access Project Final
    Report.</epdcx:valueString>
  </epdcx:statement>
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/language"
    epdcx:vesURI="http://purl.org/dc/terms/RFC3066">
    <epdcx:valueString>en</epdcx:valueString>
  </epdcx:statement>
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/type"
    epdcx:valueURI="http://purl.org/eprint/type/Report"
    epdcx:vesURI="http://purl.org/eprint/terms/Type" />
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/terms/available">
    <epdcx:valueString epdcx:sesURI="http://purl.org/dc/terms/W3CDTF">2007-08-
    17</epdcx:valueString>
  </epdcx:statement>
  <epdcx:statement epdcx:propertyURI="http://purl.org/eprint/terms/status"
    epdcx:valueURI="http://purl.org/eprint/status/NonPeerReviewed"
    epdcx:vesURI="http://purl.org/eprint/terms/Status" />
  <epdcx:statement epdcx:propertyURI="http://purl.org/eprint/terms/copyrightHolder">
    <epdcx:valueString>David Shotton, Jun Zhao, Graham klyne and Julie
    Allinson</epdcx:valueString>
  </epdcx:statement>
  <epdcx:statement epdcx:propertyURI="http://purl.org/eprint/terms/isManifestedAs"
    epdcx:valueRef="dia-final-report-1-1" />
  <epdcx:statement epdcx:propertyURI="http://purl.org/eprint/terms/isManifestedAs"
    epdcx:valueRef="dia-final-report-1-2" />
</epdcx:description>
<epdcx:description epdcx:resourceId="dia-final-report-1-1">
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/type"
    epdcx:valueURI="http://purl.org/eprint/entityType/Manifestation" />
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/format"
    epdcx:vesURI="http://purl.org/dc/terms/IMT">
    <epdcx:valueString>application/pdf</epdcx:valueString>
  </epdcx:statement>
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/publisher">
    <epdcx:valueString>Image Bioinformatics Research Group, Department of Zoology,
    University of Oxford</epdcx:valueString>
  </epdcx:statement>
  <epdcx:statement epdcx:propertyURI="http://purl.org/eprint/terms/isAvailableAs"
    epdcx:valueURI="http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres/defin
    ing_image_access.aspx" />
  <epdcx:statement epdcx:propertyURI="http://purl.org/eprint/terms/isAvailableAs"
    epdcx:valueURI="http://ora.ouls.ox.ac.uk" />

```

```
</epdcx:description>
<epdcx:description epdcx:resourceId="dia-final-report-1-2">
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/type"
epdcx:valueURI="http://purl.org/eprint/entityType/Manifestation" />
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/format"
epdcx:valueURI="http://purl.org/dc/terms/IMT">
    <epdcx:valueString>text/html</epdcx:valueString>
  </epdcx:statement>
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/publisher">
    <epdcx:valueString>Image Bioinformatics Research Group, Department of Zoology,
    University of Oxford</epdcx:valueString>
  </epdcx:statement>
  <epdcx:statement epdcx:propertyURI="http://purl.org/eprint/terms/isAvailableAs"
epdcx:valueURI="http://imageweb.zoo.ox.ac.uk/pub/2007/DefiningImageAccess/FinalReport
/DIA_FinalReport_2007-08-17.html" />
</epdcx:description>
<epdcx:description
epdcx:resourceURI="http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres/defining
_image_access.aspx">
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/type"
epdcx:valueURI="http://purl.org/eprint/entityType/Copy" />
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/terms/license"
epdcx:valueURI="http://creativecommons.org/licenses/by-nc-sa/3.0/" />
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/terms/accessRights"
epdcx:valueURI="http://purl.org/eprint/accessRights/OpenAccess" />
</epdcx:description>
<epdcx:description epdcx:resourceURI="http://ora.ouls.ox.ac.uk">
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/type"
epdcx:valueURI="http://purl.org/eprint/entityType/Copy" />
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/terms/license"
epdcx:valueURI="http://creativecommons.org/licenses/by-nc-sa/3.0/" />
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/terms/accessRights"
epdcx:valueURI="http://purl.org/eprint/accessRights/OpenAccess" />
</epdcx:description>
<epdcx:description
epdcx:resourceURI="http://imageweb.zoo.ox.ac.uk/pub/2007/DefiningImageAccess/FinalReport/DI
A_FinalReport_2007-08-17.html">
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/elements/1.1/type"
epdcx:valueURI="http://purl.org/eprint/entityType/Copy" />
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/terms/license"
epdcx:valueURI="http://creativecommons.org/licenses/by-nc-sa/3.0/" />
  <epdcx:statement epdcx:propertyURI="http://purl.org/dc/terms/accessRights"
epdcx:valueURI="http://purl.org/eprint/accessRights/OpenAccess" />
</epdcx:description>
</epdcx:descriptionSet>
```