



JISC
Digital Repositories Conference:
Dealing with the Data Deluge



University of Manchester
5-6 June 2007

Defining image access

David Shotton

oerc



Image BioInformatics Research Group
Oxford e-Research Centre and
Department of Zoology
University of Oxford, UK

e-mail: david.shotton@zoo.ox.ac.uk

© David Shotton, 2007



JISC Defining Image Access Project

Defining Image Access:

Requirements for interoperable discovery and delivery of image data stored in DSpace, EPrints and Fedora institutional repositories using a data web approach

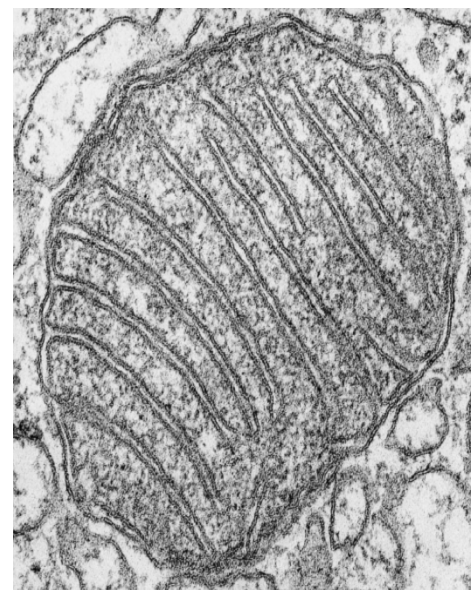
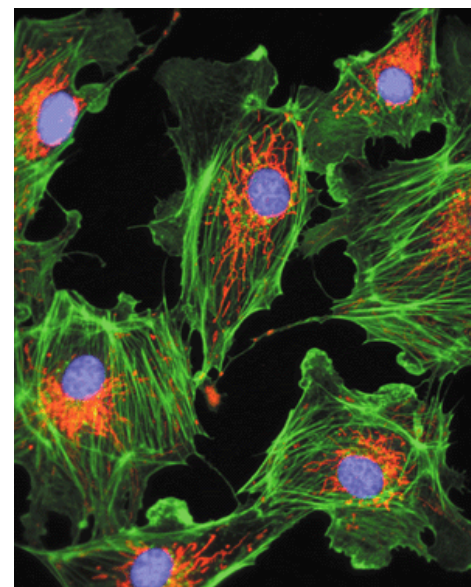
- A six-month JISC-funded requirements analysis project, Jan-June 2007
- It involves my **Image Bioinformatics Research Group** with the following partners
 - **Repository partners** at Cambridge, Imperial College, Oxford and Southampton
 - The **Oxford e-Research Centre** and **Oxford University Computing Service**
 - **UKOLN** Digital Repositories Programme Support Team
 - **CCLRC** e-Science Centre
- Looking at images and videos across discipline boundaries, not just biology
- Its primary deliverable will be a **published report** detailing our findings and conclusions, that will inform future activities
- Another significant deliverable is our **wiki**, that documents the project:
http://imageweb.zoo.ox.ac.uk/wiki/index.php/Defining_Image_Access

Outline

- The nature of scientific image data
 - What data do we actually publish?
- Integrating data from distributed resources
 - Data webs
 - ImageWeb
- The JISC *Defining Image Access* Project
 - Main achievements and conclusions
 - Proposed future activities
- Interactions with other JISC and international activities

So, what's so special about images?

- Despite popular misconception, the central problem with images is not their size or complexity
- It is that, unlike text documents, **images are not self-describing**
 - While images may be easily interpretable by humans, they are not readily amenable to automatic interpretation by present technologies
- Since their **internal semantics** are not easily extractable, descriptive **metadata annotations** are usually required to bridge this '**semantic gap**'
 - Without such metadata, digital images are virtually meaningless



What data *do* we publish?

- A scientific paper does not just report scientific observations
- Rather, as **Anita de Waard** of Elsevier has pointed out, a scientific paper is **an exercise in rhetoric**, designed to convince readers of the truth of a particular scientific hypothesis or belief
 - The goal of the article is not to state facts, but rather to **convince**
 - Facts are **selected** to support the argument, and are embedded in a rhetorical structure with the purpose of conviction

MUSCLE & NERVE 9:501–514 1986

QUANTITATIVE FREEZE–FRACTURE STUDIES OF MEMBRANE CHANGES IN CHICKEN MUSCULAR DYSTROPHY

BARBARA McLEAN, PhD, LAILA MAZEN-LYNCH, PhD,
and DAVID M. SHOTTON, PhD

“These observations support theories that defects of the muscle plasma membrane are important for dystrophic pathogenesis.”

... but what about the original research data?

- While selected findings that support hypotheses appear in research articles, the majority of original research data are **never published**
- Historically, in the paper age, there was no easy method for doing this
 - Journals had limited space
 - Other publication avenues were not available
- Now, in this digital age, data can be put on-line as 'supplementary information' on publishers' Web sites, or in institutional repositories
 - However, this facility is **not yet widely used**
 - Furthermore, such data are usually **poorly structured**, with **insufficient metadata**, and may not be discoverable by search engines
 - Depositing data in institutional repositories or elsewhere in this way may thus be consigning them to costly **data graveyards**, from which resurrection is difficult, if not impossible

How might we improve on this situation?

(my take home messages !!)

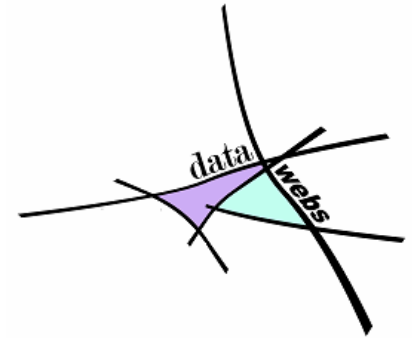
- We need to start treating experimental research data sets as **first class publication objects**, of equal value to the journal papers based upon them
 - This includes recognition and reward for data publication
- We need to ensure they are saved with **good domain-specific metadata**
 - This includes assisting researchers to capture metadata at the time of data creation, automatically if possible
- We need to work towards **better interoperability between papers and data**

Integrating distributed data

- The problems of achieving semantic interoperability between distributed heterogeneous archives of digital data and papers are well known
- Previous approaches to solving the problem have involved
 - distributed query processing
 - repository federation, or
 - portals
- All shared in common reliance on mainstream technologies such as Z39.50, XML and Web Services, some of which might be considered as dated or heavyweight technologies
- None have applied to the problems of data integration the Semantic Web / Web 2.0 approaches that I wish now to describe

Data integration - the lightweight data web approach

The data web is a novel concept for digital information integration involving semantic web technologies



- The data are held locally, with metadata published on local Web servers
- *Separately for each data web serving a particular knowledge domain,* automated **lightweight software tools** will be used to integrate the distributed data
 - separate metadata schemas will be mapped to a core **ontology**
 - **instance metadata** describing the distributed data will be made available for harvesting as RDF by creating a **SPARQL endpoint** at each resource
- Resources can then be discovered by **distributed SPARQL queries** across the data web
- This overcomes syntactic and semantic differences between data providers

Role of the data web

- The data web **aggregator**, with its associated **schema registry** and **co-reference service**, acts first as a data marshal, ordering and integrating the metadata schemas from the data web participants into a single searchable RDF graph
- It then provides a **SPARQL query handler**, providing access to all the data in the data web, with both human and programmatic access
 - Remember: *With RDF, integration comes for free!*
- **The data web adds value** by providing interoperability and customizable search interfaces, **with a rigorous semantic underpinning**
- **The primary data holders benefit** by **increased user traffic to their sites**, while being able to **maintain normal copyright and access control**
- **The primary metadata** are **never controlled by the data web**, but are freely available on the Web for use by other presently unforeseen applications, including novel data mining, integration and analysis services

The ImageWeb Project



- **Image webs are data webs for research images**
- We desire to integrate and make cross-searchable **research images** held by **publishers**, **research organizations**, **museums and learned societies**, and **institutional repositories**, which are currently in isolated data silos
- We desire to enable these information resources
 - to become a **more integral part of day-to-day research**, and
 - for published images to be **more fully used** than at present, including combination and re-use for **meta-research**
- The same images might be accessed by more than one data web
 - For example, cellular images might be accessed by one data web illustrating **confocal microscopy techniques**, and alternatively by another data web concerned with **cancer therapy**

ImageWeb

The BioImageWeb Consortium

- Image BioInformatics Research Group, University of Oxford



- Leading commercial publishers
 - Nature, OUP, Blackwell, Elsevier

nature

OXFORD
UNIVERSITY PRESS

Blackwell
Publishing

ELSEVIER

- Leading Open Access publishers
 - The Public Library of Science and BioMed Central

PLOS

BioMed Central
The Open Access Publisher

- University institutional repositories
 - Universities of Cambridge, Imperial College, Oxford and Southampton

 **UNIVERSITY OF**
CAMBRIDGE

Imperial College
London

University of Oxford

 **University**
of Southampton

- Other stakeholders: BL, CCLRC, UKOLN, ILRT, CrossRef, SPARC, Ingenta

BRITISH
LIBRARY

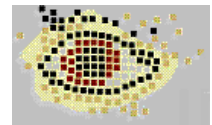
 **CCLRC**

 **UKOLN**

ILRT

crossref

 **SPARC**
Europe



- Professional researchers and academic image collections

 **NATURAL**
HISTORY
MUSEUM

Stakeholder analysis

Stakeholder	Interest / stake
Research funding agencies	Full publication of, and access to, research results from funded projects Reduce future research costs by facilitating re-using of hard-won observational data
Researchers creating data, and requiring access to others' data	Publish their own research data more easily Locate on-line data (including images) easily, with sufficient supporting metadata to permit proper interpretation Follow new lines of <i>in silico</i> research based on published observations
Institutional repository managers	Serve the needs of research users effectively and economically Long-term preservation of research publications and data Enhance institutional visibility by improving access to resources
Academic publishers	Adapt successfully to changing publication and access models Provide improved access to data and images in journal articles Develop secondary added-value services Maintain normal access and copyright controls

Making the image web vision a reality



- **Phase One** (Late 2005)
 - Developing initial concepts
- **Phase Two** (2006)
 - Bringing together BioImageWeb Consortium partners
 - Organizing and running the Research Information Network **Data Webs Workshop** (<http://www.rin.ac.uk/data-webs>)
 - Applying for funding to start research work
- **Phase Three** (Jan - June 2007)
 - **Requirements analyses**
 - JISC-funded *Defining Image Access* Project for repositories**
- **Phase Four** (Autumn 2007 - Spring 2009)
 - Creation of a demonstrator - a real data web for real research images
- **Phase Five** (Future)
 - Routine use of data webs in support of scholarly research

JISC *Defining Image Access* Project

Requirements Analysis Activities

- Analysis of current **institutional repository practice** with particular reference to image storage
- Evaluation of **repository software capabilities**, particularly with respect to
 - serving domain-specific metadata
 - metadata harvesting using OAI-PMH
 - programmatic access
- Evaluation of third-party **Semantic Web approaches and software tools** to fulfil the functional requirements of a data web
- Evolution of our **concepts** of what a data web for images might entail, driven by potential uses of image webs in other research projects
- Full details of all our interviews and evaluations are on our Wiki:
http://imageweb.zoo.ox.ac.uk/wiki/index.php/Defining_Image_Access

JISC Defining Image Access Project

Achievements

- We have established a significant core body of **knowledge and expertise** concerning images in institutional repositories, and the problems and opportunities associated with integrating them, that is freely available for everyone to access from our wiki
- We have held three very productive **project workshops**, that have served both to permit us to learn from other projects and also to publicize what we are doing
- We have met some excellent **people** involved in **other JISC projects**, and are learning how best to integrate our activities with theirs
- We have significantly refined our ideas about **data web functionality**
- We have seen significant take-up of our data web idea for use in proposed **new projects** in the arts and the sciences, for which grant applications made
- Our final project meeting will be in Oxford on **Friday 22nd June**
- Anyone interested in coming please e-mail graham.klyne@zoo.ox.ac.uk

JISC Defining Image Access Project

Main Conclusions - Data Web Technology

- Semantic Web/Web 2.0 approaches and tools *can* be used to build data webs
- Data webs should comprise *independent loosely coupled services*
- The central data web *aggregator*, which will work with the schema registry and co-reference service to act as a query handler, *needs to be created*
- The degree to which it will be useful for the aggregator to *pre-harvest core metadata* into a central metadata registry is a subject for research
- The schema registry and co-reference service will require *hand-crafted alignment* for each knowledge domain
- *mSpace* (<http://www.mspace.fm/>) seems well suited to provide semantically enabled browse and search discovery services
- We wish to incorporate user annotation facilities, possibly using RichTags
- Building a data web remains a *significant* implementation task

JISC Defining Image Access Project

Main conclusions - Repositories

- Institutional repositories currently contain **few image collections**
- Existing image collections mostly **lack adequate domain-specific metadata**
- Repository software is **not equipped** to serve domain-specific metadata, even if it existed
 - **ePrints**: We are exploring with **Chris Gutteridge** (ePrints Soton) the possibility of adding two extra metadata fields to basic ePrints OAI-PMH DC profile, one providing the URI for a separate file containing domain-specific metadata and the other specifying its format
 - **Fedora**: Total flexibility, but you have to build your own interface !
 - **DSpace**: Serves basic OAI-PMH Dublin Core only
- These limitations make the creation of domain-specific inter-repository image webs both technically difficult and functionally unimportant

Our proposed future activities

- To build a demonstration image web linking existing publications in local institutional repositories with existing image datasets housed elsewhere
- To implement this using a defined biological domain in which digital images form the primary evidence for research conclusions
- To evaluate the usefulness of such an approach with real researchers

Additionally, in a related project to provide semantic enhancement of the scholarly lifecycle in biology:

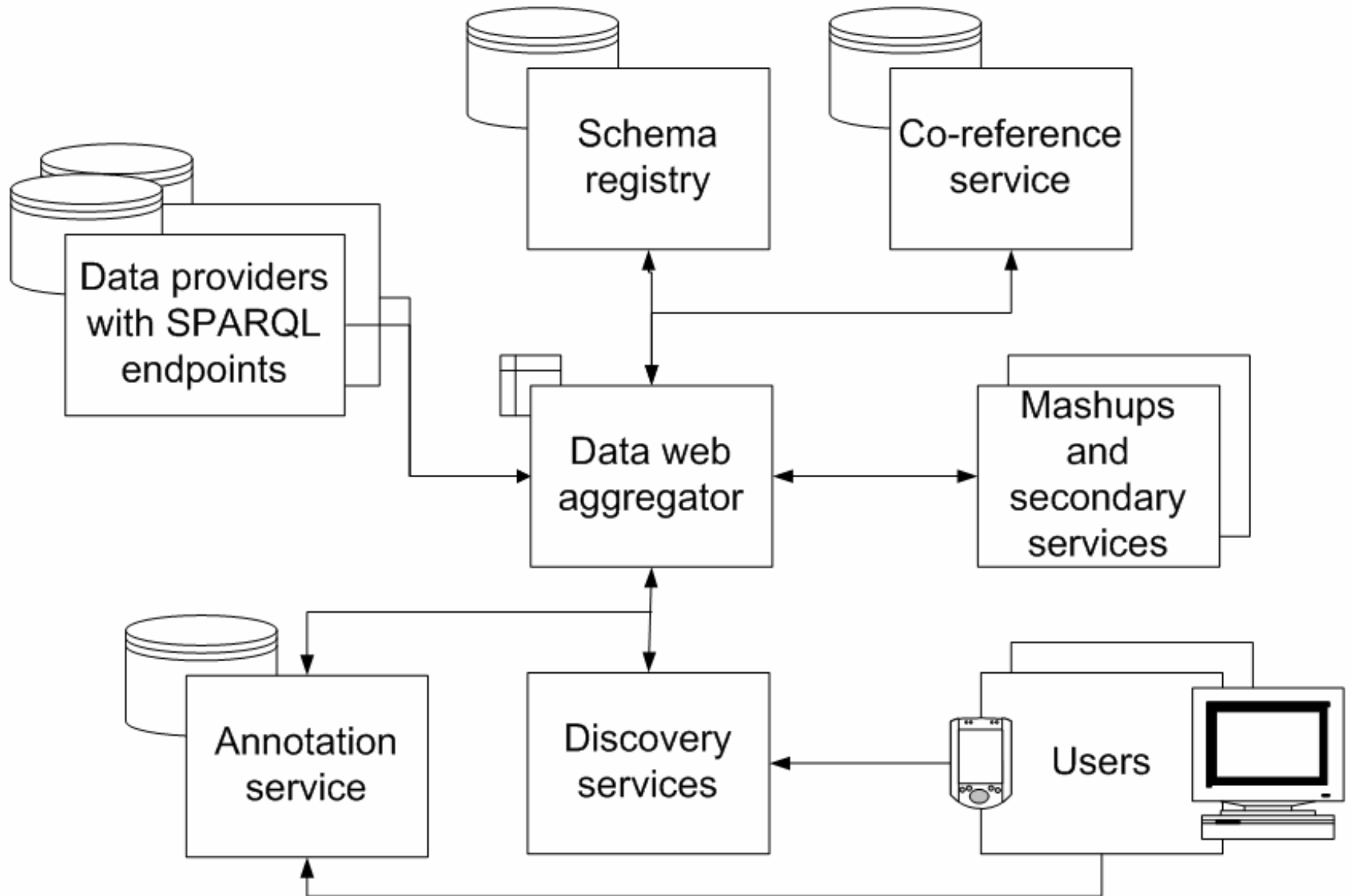
- To install a local ePrints server in our own lab, to act as an image database for research images, either private or public/published
- To facilitate (make easy!) researchers' data submissions of experimental images and videos to this database, with rich domain-specific metadata
- To facilitate migration of data from this database to institutional repositories (ePrints and Fedora initially) for long-term preservation
- To work with publishers (Wiley-Blackwell, Elsevier) to put links to datasets into on-line papers

Technical approaches

- Use of **the Web as the platform** - i.e. data webs as sets of web services
- Use of World Wide Web Consortium Semantic Web tools and standards:
 - **RDF** as the standard format for sharable metadata
 - **SPARQL** as the universal query language for RDF
 - Software such as **D2R Server** for creating SPARQL endpoints on relational databases, enabling extraction of data as RDF in response to SPARQL queries
 - **OWL-DL** as the standard web ontology language
- and for software development and integration:
 - use of agile programming techniques
 - **Ruby** or **Python** to provide a lightweight development environment
 - loose coupling between the Model, View and Controller software components, based on a simple '**REST**'-full approach to component integration ([Fielding 2000, Representational State Transfer](#))



Data web services



Interactions with other's activities

mSpace and RichTags

- **mSpace** (<http://www.mspace.fm>) is a data presentation system developed by **monica schraefel** that provides a semantically aware **faceted browse service** over semantically rich data available as RDF
 - We believe this would form an ideal user-oriented discovery service for data webs
- **monica schraefel** is currently developing a semantically rich **user annotation service** called **RichTags** in a current JISC Repositories and Preservation project
 - We plan to collaborate with her to implement RichTags as the **annotation service for data webs**

Interactions with other JISC activities

JISC Common Repositories Interfaces Working Group

- The **JISC Common Repositories Interfaces Working Group** is currently “addressing priorities for enhancing inter-working between the emerging base of UK repositories”
- I have emphasised to this group the **need for the various repository software systems to present SPARQL endpoints**, so that their exposed metadata might be easily usable by Semantic Web applications that use RDF as their native data format
- Since the Working Group has not been funded to *implement* their own recommendations, I would like their endorsement for our own JISC application for funding to create **generic SPARQL end-points for ePrints and Fedora** repositories, as part of our ongoing activity
- This functionality is integral to our plans to deploy a demonstrator data web across institutional repositories and other data sources

Interactions with other JISC activities

JISC Metadata Schema Registry

(<http://www.ukoln.ac.uk/projects/iemsr/>)

- Its activities so far, with publication of its excellent **ePrints Scholarly Works Application Profile**, and with an **Application Profile for Images** in the offing, have focused on basic Dublin Core-type metadata
- Our own proposed **Data Web Schema Registry** service is quite different, involving hand-crafted alignment of complex schemas of participants in each data web, over a fuller range of generic and domain-specific metadata
- We need to have further discussions with **Emma Tonkin** and **Julie Allinson** to see whether the JISC Metadata Schema Registry might be expanded to undertake this role as a generic service for data webs
- We are also keen to interact with **Karla Youngs** and her colleagues at TASI to provide input for the Images Application Profile

Interactions with other JISC activities

Intute Repository Search

- We are in continuing discussions with **Caroline Williams** (Director) and **Phil Cross** (Developer) about how best to **integrate** data webs with the planned Intute Repository Search
- While details have yet to be worked out, we are hoping that Intute Repository Search will be able to use **programmatic access to data webs** to enrich its own search capabilities

Interactions with other JISC activities

eBank UK, R4L and SPECTRa

- The JISC-funded **eBank UK, R4L and SPECTRa** projects have set the standard for the semantic integration of research data with repositories and scholarly publications, operating in the specific areas of chemistry and chemical crystallography
- Those domains are characterised by a unified data notation and defined conventions
- In contrast, the biological research domain is characterised by **highly heterogeneous** research data, often with little by way of semantic mark-up
- Nevertheless, there is much more we can still learn from **Simon Coles, Alan Tonge** and their colleagues concerning these excellent projects, as we seek to deploy data webs

Interactions with other JISC activities

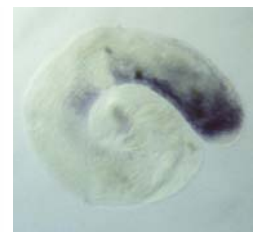
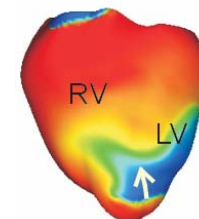
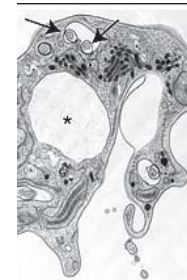
JISC-funded Digital Curation Centre SCARP Project

- Our proposed data web activities will benefit from our **current work** as part of the JISC-funded Digital Curation Centre SCARP Project on the

ImageStore Project: Curation requirements for preservation of legacy analogue and 'born digital' scientific image data in institutional repositories

ImageStore

- We are undertaking requirements analysis, investigating the current practice and future requirements for the preservation of biological research images derived from four distinct areas of research:
 - Two sets of historical analogue records
 - Electron micrographs of microtubules in trypanosomes
 - Wildlife videos of the behaviour of badgers and foxes
 - Two sets of modern 'born digital' images
 - Computer simulations of the human heart
 - Gene expression in *Drosophila melanogaster*



Interactions with other JISC activities

StORe and CLADDIER

- The **StORe** and **CLADDIER** Projects have both undertaken requirements analyses for linking repository publications to datasets housed elsewhere
- StORe has revealed the desirability of such interoperability
- CLADDIER has built an **on-line data and publication discovery service** that works well (<http://isegserv.itd.rl.ac.uk/claddier/search/single/>)
 - While these projects have not employed Semantic Web technologies, we have much to learn from them
- We propose to involve **Alistair Miles**
 - co-author of **SKOS** (the Simple Knowledge Organisation System)
 - creator of the **CLADDIER discovery service**
 - who is currently working with us on the SCARP **ImageStore Project** as a partner in our future work

Conclusion: our future objective

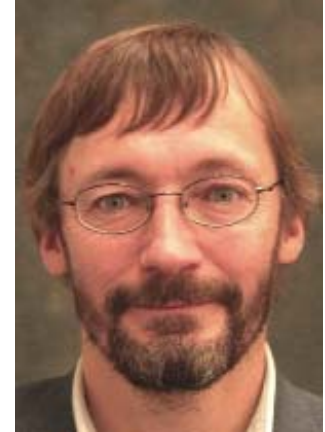


DW-40 : To create a demonstration data web providing frictionless interoperability between scientific publications and research datasets

The end

Acknowledgements - special thanks to:

Graham Klyne, with whom my data web ideas have been developed, and who has been Project Manager for the *Defining Image Access* Project



Jun Zhao, who has recently joined my group specifically to work on the *Defining Image Access* Project



JISC for funding this work, and in particular to Balviar Notay, our JISC Programme Manager



Interactions with international activities

OAI-ORE Project - Object Reuse and Exchange

- There is much discussion of the **ORE Project** (<http://www.openarchives.org/ore/>), funded by the Mellon Foundation in the USA, and led by **Carl Lagoze** and **Herbert Van de Sompel**
- It is a successor to the Pathways Project to develop a loosely-coupled, highly distributed, interoperable scholarly communication system
- ORE will develop **specifications** that allow distributed repositories to exchange information about their constituent digital objects
- Others expect it to have an impact as significant as that of OAI-PMH
- However, it does NOT propose to create implementations of these specifications, and it does NOT use Semantic Web technologies
- Both the chair of the JISC Common Repositories Interfaces Working Group and our *Defining Image Access* partner UKOLN represent the JISC on this project, and we are looking to them to keep us informed about OAI-ORE

Image webs are of generic applicability

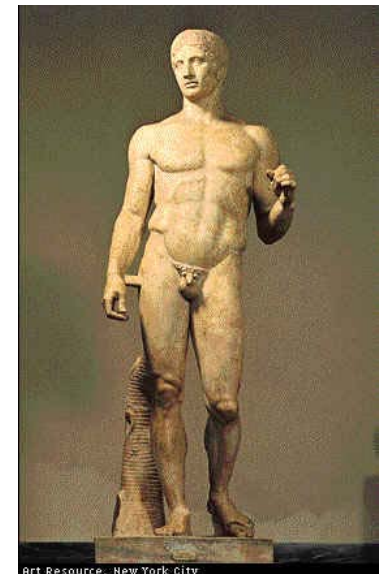
- **CoKE Project - Collaborative Knowledge Extraction : Sequencing Animal Behaviour - a RCUK Basic Technology Programme application**

- Use of data web technology and our Animal Behaviour Ontology to assist in the high throughput screening of genetic defects or of drug effects in model organisms

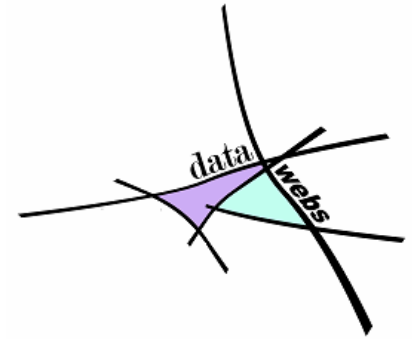


- **myEvent Project - an EC FP7 application**

- Metadata and images of classical art museum objects linked by an image web
- Multi-lingual descriptions at different levels of complexity (child, teenager, adult, scholar) provided to museum visitors via mobile phones
- 3D virtual displays and supplementary information, and the ability to create virtual **exhibitions@home**



How might a data web improve on Google™?



- It permits access to database information hidden in the 'Deep Web'
- It involves **specific targeting to a particular knowledge domain**, thus achieving a significantly higher signal-to-noise ratio
- It provides **integration of information** with ontological underpinning, semantic coherence, and truth propagation
- It permits **programmatic access**, enabling **secondary services** to be built on top of one or more data webs

Web 2.0 aspects of data webs



- Use of the Web as the platform
- Small pieces, loosely coupled
- Programmatic access, giving 'hackability' and the right to remix
- Tagging:
 - Data webs are predicated on a formal core ontology, but we see vital roles for user annotations to supplement formal metadata
- Trusting our users:
 - Data providers control their own primary image data and metadata
 - Data consumers are free to use the data web service in whatever way they think fit, including building secondary services, and providing annotations
- The Long Tail:
 - Data webs enable discovery of 'long tails' of hard-to-find data - this is particularly true for 'research particulars' such as images