

Large-scale data sharing in the life sciences: Data standards, incentives, barriers and funding models (The “Joint Data Standards Study”)

Prepared for

The Biotechnology and Biological Sciences Research Council
The Department of Trade and Industry
The Joint Information Systems Committee for Support for Research
The Medical Research Council
The Natural Environment Research Council
The Wellcome Trust

Prepared by

The Digital Archiving Consultancy (DAC)
The Bioinformatics Research Centre, University of Glasgow (BRC)
The National e-Science Centre (NeSC)

August 2005

Table of contents

VOLUME 1 - Report	Page
Abbreviations	4
Acknowledgements	5
Executive summary	8
PART 1	
1. Background to report	13
2. Method	14
3. Key concepts	17
PART 2	
Findings: general introduction	20
4. Technical issues	21
5. Tools	27
6. Standards	33
7. Planning and management over the data life cycle	37
8. Nature of content	45
9. Organisational and management issues	50
10. Culture and careers	55
11. Support, training and awareness	59
12. Legal, regulatory, and rights management issues	61
13. Corporate sector	64
14. International issues	68
PART 3	
15. Core models for data sharing	71

VOLUME 2 – Appendices	Page
Appendix 1: Bibliography	A-2
Appendix 2: Glossary	A-10
Appendix 3: Case studies	A-19
Case study 1: Arts & Humanities Data Service	A-23
Case study 2: BADC / NERC DataGrid	A-29
Case study 3: BRIDGES	A-38
Case study 4: CLEF	A-47
Case study 5: Common Data Access	A-54
Case study 6: Ensembl	A-61
Case study 7: Genome Information Management System	A-67
Case study 8: Malaria (<i>Plasmodium falciparum</i>)	A-72
Case study 9: Nottingham Arabidopsis Stock Centre	A-76
Case study 10: Proteomics Standards Initiative	A-83
Appendix 4: Further sharing models	A-97
Appendix 5: Supplementary materials	A-104
5.1 The data provenance issue	A-104
5.2 Standards and data sharing	A-109
5.3 Databases	A-113

Abbreviations commonly used in the text

The following abbreviations are used commonly throughout the text. Less frequently used abbreviations are listed in Appendix 2.

AHDS	Arts and Humanities Data Service
AHRC	Arts and Humanities Research Council
BADC	British Atmospheric Data Centre
BBSRC	Biotechnology and Biological Sciences Research Council
BGS	British Geological Survey
BODC	British Oceanographic Data Centre
BRIDGES	Biomedical Research Informatics Delivered by Grid Enabled Services
CDA	Common Data Access
CLEF	Clinical eScience Framework
DCC	Digital Curation Centre
DTI	Department of Trade and Industry
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
GIMS	Genome Information Management System
ICT	Information and Communications Technology
IP	Intellectual property
IT	Information technology
JCSR	JISC Committee for the Support of Research
JISC	Joint Information Systems committee
MIPS	Munich Information Centre for Protein Sequences
MRC	Medical Research Council
NASC	Nottingham Arabidopsis Stock Centre
NCRI	National Cancer Research Institute
NDG	NERC DataGrid
NERC	Natural Environment Research Council
NHS	National Health Service
NIH	National Institutes of Health
NTRAC	National Translational Cancer Research Network
OECD	Organisation for Economic Cooperation and Development
PDB	Protein Data Bank
PI	Principal investigator
PSI	Proteomics Standards Initiative
W3C	World-Wide Web Consortium
XML	eXtensible Mark-up Language

Acknowledgements

We would like to thank the following for their generous and substantial contribution to the preparation of this report, as interviewees for case studies or as key informants. (In brackets we indicate the name of a case study, where a meeting was specifically related to a case study or a specific project.)

Dr. Sheila Anderson	Arts & Humanities Data Service (AHDS)
Dr. Rolf Apweiler	European Bioinformatics Institute
Professor Geoff Barton	University of Dundee
Dr. Andrew Berry	Wellcome Trust Sanger Institute (Malaria)
Dr. Ewan Birney	European Bioinformatics Institute (Ensembl)
Ian Bishop	Foster Wheeler Energy UK
Professor Andy Brass	University of Manchester
Professor Graham Cameron	European Bioinformatics Institute
Tara Camm	The Wellcome Trust
Professor Rory Collins	University of Oxford
Dr. Michael Cornell	University of Manchester (GIMS)
Damian Counsell	Rosalind Franklin Centre for Genomic Research
David Craigon	NASC University of Nottingham (NASC)
Dr. Richard Durbin	Wellcome Trust Sanger Institute (Ensembl)
Dr. Dawn Field	NERC Environmental Genomics Programme
Dr. Andrew Finney	University of Hertfordshire
Malcolm Fleming	CDA Limited (CDA)
Professor Jeremy Frey	University of Southampton (eBank)
Kevin Garwood	University of Manchester
Dr. Peter Ghazal	University of Edinburgh
Dr. Neil Hanlon	University of Glasgow BRC (BRIDGES)
Dr. Rachel Heery	UKOLN (eBank)
Dr. Kim Henrick	European Bioinformatics Institute
Professor Pawel Herzyk	University of Glasgow (BRIDGES)
Derek Houghton	University of Glasgow, BRC (BRIDGES)
Dr. Tim Hubbard	Sanger Institute (Ensembl)
Professor Michael Hursthouse	University of Southampton (eBank)
Professor David Ingram	University College, London (CLEF and OpenEHR)
Dr. Hamish James	Arts & Humanities Data Service (AHDS)
Dr. Nick James	NASC University of Nottingham (NASC)
Dr. Dipak Kalra	University College London (CLEF)
Dr. Arek Kasprzyk	European Bioinformatics Institute (Ensembl)
Professor Douglas Kell	University of Manchester
Dr. Peter Kerr	National Cancer Research Institute
Dr. Bryan Lawrence	Rutherford Appleton Laboratory (BADC/NDG)
Professor Ottoline Leyser	University of York
Dr. Liz Lyon	UKOLN/DCC (eBank)

Gary Magill	Foster Wheeler Energy UK
Dr. Bob Mann	Institute for Astronomy, University of Edinburgh & DCC
Dr. Paul Matthews	European Bioinformatics Institute
Dr. Sean May	NASC University of Nottingham (NASC)
John McInnes	British Geological Survey (CDA)
Dr. Muriel Mewissen	University of Edinburgh
Helen Munn	Academy of Medical Sciences
Professor Christopher Newbold	University of Oxford (Malaria)
Professor Steve Oliver	University of Manchester (GIMS)
Professor Ken Peach	The Council for the Central Laboratory of the Research Councils
Dr. Chris Peacock	Wellcome Trust Sanger Institute
Dr. Robert Proctor	University of Edinburgh (NTRAC)
Professor Alan Rector	University of Manchester (CLEF)
Dr. Fiona Reddington	National Cancer Research Institute
Dr. Peter Rice	European Bioinformatics Institute
Beatrice Schildknecht	NASC University of Nottingham (NASC)
Steven Searle	Wellcome Trust Sanger Institute (Ensembl)
Peter Singleton	Cambridge Health Informatics (CLEF)
Dr. Barbara Skene	The Wellcome Trust
Dr. Arne Stabenau	European Bioinformatics Institute (Ensembl)
Dr. James Stalker	Wellcome Trust Sanger Institute (Ensembl)
Professor Alex Szalay	Dept. Physics, John Hopkins University, USA, and project director of US National Virtual Observatory
Dr. Chris Taylor	European Bioinformatics Institute (PSI)
Dr. Bela Tiwari	NERC Environmental Genomics Programme
Dr. Alexander Voss	University of Edinburgh (NTRAC)
Professor Michael Wadsworth	University College, London
Martin Wadsworth	CDA Limited (CDA)
Dr. Matthew West	Shell UK
Dr. Anne Westcott	AstraZeneca
Dr. Max Wilkinson	National Cancer Research Institute
Dr. Emma Wigmore	NASC University of Nottingham (NASC)
Dr. Michael Williamson	University of Manchester
Professor Keith Wilson	University of York
Dr. Matthew Woollard	Arts & Humanities Data Service (AHDS)

In addition to these interviews, several events and other projects provided the authors with opportunities to talk with many experts and practitioners in the life sciences, environmental and geophysical sciences, and IT.

The 2004 e-Science All Hands Meeting was an opportunity to hear presentations about many areas and projects of relevance to this report, and to talk to many key informants. Our thanks in particular to Professors Carole Goble and Norman Paton of the University of Manchester.

The 2004 International Systems Biology Conference presented an extended opportunity to talk with key informants from all around the world: from the UK, Professor Andy Brass of the University of Manchester, Dr. Albert Burger, Heriot-Watt University, and many from further afield, from America, France, Germany, South Africa. Of the many relevant presentations and posters, there was a very interesting debate centred around a discussion of strengths, weaknesses, opportunities and threats, at the Ontologies Special Interest Group chaired by Helen Parkinson (EBI), with Dr. Jeremy Rogers (University of Manchester, CLEF), Dr. Crispin Miller (University of Manchester), Professor Barry Smith (University of Buffalo), and Professor Michael Ashburner, and a helpful day on modelling in systems biology, chaired by Professor Igor Goryanin and Dr. Ian Humphery-Smith.

The JISC runs many projects and studies relevant to this study; the JISC July 2004 Conference included presentations by Alan Robiette (JISC), Jane Barton on a JISC study on metadata, Brian Kelly on quality assurance for libraries, and Jon Duke and Andy Jordan on their study on Developing Technical Standards and Specifications for JISC Development Activity and Service Interoperability.

The clinical domain presents particular problems in data sharing. Our work has been enormously helped by involvement in the Consent and Confidentiality project conducted by the MRC. Several events helped inform this study, including a one-day symposium was held in October 2004, at the Royal College of Physicians in London, on the Governance of Medical Research Databases, and a one-day meeting on the e-Bank project in August 2004 at UKOLN.

Above all, however, the authors would like to thank the sponsors for the opportunity to work on this extremely interesting project. We thank the team at the Medical Research Council, for their guidance, support and help, in particular Dr. Peter Dukes, Dr. Allan Sudlow, Sharon Etienne and Nicola Wiseman. All the sponsors' representatives have been most helpful in providing suggestions and guidance.

EXECUTIVE SUMMARY

Recent studies have demonstrated the value of sharing and re-using data in the life sciences. This study builds on that premise, exploring the practice of sharing data and identifying incentives, facilitators and obstacles to data sharing. In the light of its findings the study presents characteristics of models which support effective, ethical data sharing, to enable first-class, innovative and productive science, within and across disciplines.

The sponsors' objectives are that maximum and optimum use be made of expensively generated and expensively maintained data. Indeed, the study found that difficulties in sharing data have a substantial cost, above all in wasted time. Some quantification is provided in corporate figures which estimated data exchange or retrieval costs at 3% of a project's capital cost or up to 60% of an engineer's time.

The study spanned the full spectrum of the life sciences, from genome to ecosystems, and all data types and methods, from bench science to field surveys, including data-based science. It examined ten case studies, including two comparators from outside the life sciences. These case studies were supplemented by interviews with key informants and by desk research.

Data sharing here means the re-use of data, in whatever way, wherever it is held, at any time after its creation (however long). Re-use includes actions on data such as consultation, comparison, analysis, integration, transformation, annotation, and enhancement. "Data" in this context is interpreted broadly, to include (for example) metadata and processes. Entities that supply data and associated services are referred to as "data resources" or "community resources".

Standards are fundamental to data sharing at all levels, from administrative and management, software, infrastructure, to data and scientific standards: They help overcome interoperability difficulties across different data formats, architectures, and naming conventions, and at infrastructure level, enabling access and authorization systems to work together. The study found that absence of standards means substantial loss of productivity and less data available to researchers.

Standards relating to data and information need to be sufficiently flexible that they can work across a broad spectrum and cope with the rapid pace of change at both scientific and IT levels. Information and data standards need to be flexible enough to serve a wide user base and adjust to rapid science and technology change. Successful examples seen were cast at a high level, avoiding detail, and were extensible. Ease of use was key to adoption by users. Achieving this involved considerable work, at national and international level: it needs a formal development framework, community consultation, pragmatic and disciplined management, funding for those involved in the development of standards, and continued support and funding for those implementing and encouraging their uptake.

Data sharing requires **planning and management**. A laissez-faire approach to the collection and distribution of data results in waste, notably data with insufficient information to enable re-use.

- At project level, sharing has to be anticipated at the beginning of the data life cycle, at funding application stage, whether within publicly-funded or corporate research, by forecasting a project's digital output, identifying the information that needs to be collected to support downstream sharing and the resources (tools, equipment, people's time, skills) for collecting that information. This planning should be drawn up by applicants at funding application, should be done in consultation with the funder's representative(s) and should be informed by

funder policies. The plan should be drawn up in liaison with the destination repository and/or body which catalogues data, enabling it to be located and preserved. These plans need to be reviewed and updated over the course of the project.

Ethical approval for studies to proceed needs to include permission for future data sharing and long-term preservation, the need to obtain anticipatory or generic consent from the data subjects, and the possible use of anonymisation or pseudonymisation to enable future accesses to personal data beyond the bounds of consent given.

- Data must be managed during projects and during its life after project end.
- Technical difficulties in re-using data are substantially eased by good data management (including documentation) and timely collection and provision of data about the data, i.e. metadata. Data generators must provide adequate, accurate metadata to their digital output at time of creation for efficient, cost-effective, quality data sharing downstream. The metadata requirements must be identified *a priori*, at scientific but also at computing and administrative levels. This requires co-ordination, including between computer and information science and scientific disciplines. Tools help the timely collection of good quality metadata.
- Planning and management needs to be informed by defined objectives, policy and strategy, and these need to take account of the needs of the full life cycle of the data from planning to archiving. The information will provide policy makers and management with increased information about scientific output, and scientific, IT and infrastructure needs. They will then be in a better position to identify priorities, needs and also economies of scale.
- For the sustainability of the data-sharing base there needs to be co-ordination between scientific data management initiatives (such as MIAME) and data/information management experts, such as the British Atmospheric Data Centre, the Arts and Humanities Data Service, the Digital Curation Centre, and UKOLN. For the richness of the knowledge base, there should be co-ordination with information centres of excellence such as the British Library, the Research Information Network, and JISC initiatives and e-Science projects (such as such as e-Bank and BRIDGES, which specifically weave together different forms and types of information, and different disciplines).

Planning and management predicate the existence of a **framework** that **supports** researchers in their data collection and submission activities. Excellent exemplars are provided by the NERC's Environmental Genomics programme, within the NERC data management programme, and the Arts and Humanities Research Council's Arts and Humanities Data Service (AHDS):

- At a practical level, support, training and guidance are provided to researchers on data management, couched in plain, accessible English. The NERC's Environmental Genomics programme draws together scientific, computing and data management expertise to anticipate tools and services to support active research and data collection. Similarly, the AHDS brings together specialist discipline and technical expertise to advise users, and it plays an active role at the forefront of work on digital preservation and curation.
- Training enables researchers to make fuller use of both data and tools: without it, data and tools either lie fallow or are under-exploited.

- Awareness campaigns showing scientific benefit through data sharing enhance use of data management support services and user engagement in the data-sharing process.

Software tools are vital for data sharing, but if they are to be of use to the scientific community they need to be developed and maintained to quality levels. Specialist informatics units usually have staff with experience of developing software to quality standards, working in frameworks where testing is carried out and documentation maintained. The result is software tools which are robust, and errors can be readily eradicated: neither documentation nor tool gets lost and the user is not subjected to time-wasting frustrations getting them to work. Users' input into tool design and ease of use are key to use of the tool. Interoperability of tools across computing platforms and across scientific domains increases their value.

Efficient use of data held in community resources relies on **good communication**, documentation and support cover available as advertised. Delivery and service to users should be the primary focus of community resources. To achieve this, community and data resources should have, with the support of their funding agencies, defined delivery objectives, with funding to match. Proximity to related expertise brings synergies and cross-fertilization of ideas. The location of data resources within a wider organisational framework facilitates efficient management over the life cycle of data resources, in particular in the case of retirement of resources whose level of use no longer requires active resource provision.

Incentives are needed for the engagement of individuals at the data creation stage, to submit data for sharing, and to supply the metadata that is needed for others to share the data downstream. In achieving this both funders and individuals are key to a sustained, open culture of sharing.

Measures such as setting periods of privileged access to data, and recognition of (and mechanisms for) data citation will help provide incentives. Sensitivity to individual circumstances will avoid hostility in the community. Data management, curation accreditation schemes should be developed to foster trust on the part of users in the quality of data. Work (underway) for accreditation for repositories and data resources will help in this regard. Issues such as assured provenance for digital data are still questions of research and refinement of practice; these are particularly important in the clinical domain.

Data sharing involves more than just exchange of data. We argue that **culture and careers** influence the life-cycle, availability and quality of data. Again, sensitivity to individual circumstances when setting periods of privileged access helps overcome resistance to or fear of data sharing.

There should be recognition of the role and skills set of data resource managers; promotion boards should be briefed on the needs for and the needs of data curation, data management and archiving skills, which often do not fit current career categories. Amongst criteria by which researchers are judged, recognition of data citation and contributions to repositories will help data sharing. This in turn will provide incentives to users to participate in data sharing actions and foster awareness and skills.

Data sharing **must respect the consent given** for data use and the confidentiality of data provided by research subjects. Failure to do so will result in withdrawal of consent by subjects and refusal to use the digital medium by practitioners. Considerable difficulties can be encountered obtaining consent, but the consent process needs to address the potential for re-use, balancing benefits and risks for the various stakeholders. This issue is complex and needs further work at technical, procedural and policy levels; it is complicated by responsibilities being dispersed across organisations, interests and

boundaries (local, national and international). The success of projects such as CLEF, BioBank, and the NHS National IT Programme, will be critical to the public confidence needed to support sharing of personal data for research.

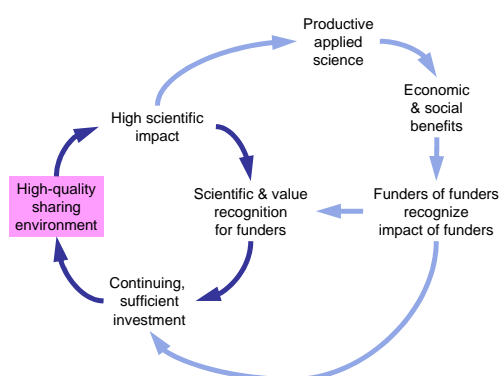
Funding: Resources that serve a wide community should be publicly funded; furthermore their maintenance, integration with other data and their further development should be funded too. Funding should be stable enough to allow medium-term planning and assured longevity. One way of achieving this is through “rolling” grants with long closure notices and, where relevant, exit plans in place. Good data management during the execution of projects also needs resource allocation, in particular the process of submitting data to a data resource. For data resources, funding is needed not only for data management, access and preservation, but also for providing adequate support to their users, whether depositors or consumers.

Legislation can act both positively for and negatively against data sharing. More research needs to be done and taken into account in informing and framing legislation which might affect the sharing or re-use of digital objects (data, software, processes).

With regard to rights management, mechanisms need to be sought for enabling “time-proof” rights to accompany data; this is akin to the provenance issue noted on the previous page.

Data sharing crosses national boundaries. It is critical to research into areas such as climate change and biodiversity. International and multilateral policy must be supported by action. Co-ordination on data sharing on an international scale is necessary.

The study shows that data sharing contributes to a virtuous circle, where promoting effective sharing widens research and enhances scientific impact. This scientific impact can be measured in the number and quality of publication and data citations, prizes, in the generation of new science and increased quality of research outcomes. These redound to funding agencies, in turn encouraging continued, sustained funding, which in turn feeds back into sharing and research.



Summary of the characteristics of effective data-sharing processes:

Based on our discussions and reading in the preparation of this report, we can outline several characteristics of effective data sharing:

1. Data and appropriate, adequate and good-quality metadata are provided by data generators based upon accepted technical standards and using standard vocabularies.

2. The resources provided are visible to potential users within and outside community boundaries, and navigation to the resources is clear and reliable.
3. As far as possible, access to data and metadata should be openly available (with due recognition of intellectual property and sources), except where non-disclosure or controls are appropriate to protect individuals, environments or species.
4. Where data cannot be made publicly available, but is shared between a given closed set of authorized users, technological solutions should help maintain privacy and security.
5. The validity and quality of the data can be established rapidly and reliably by users.
6. Support, training, and other help to those engaged in data sharing (and curation) are readily locatable and available at all points in the process, from planning to data retirement.
7. The primary focus of community resources is delivery to users, matching objectives and policy. Delivery is enriched and sustained by research.
8. Positive incentives exist for suppliers of data, and measures to reassure them that their careers and research are enhanced by participation in the data-sharing process.
9. For staff managing data resources there are appropriate career structures and rewards. Career paths should recognise the new domains of (for example) bio-informaticians and data curators.
10. Actively managed frameworks which exploit proximity of resources, expertise and economies of scale facilitate sustained and appropriate funding over the life of the resources.

Key recommendations

We propose some key recommendations which we believe should be incorporated within future data-sharing policies:

1. Insistence on a data management plan as part of the funding application and evaluation process; police the spend; provide incentives to spend the data management allocations on data management. (Recommendation 7.1)
2. Resources should have a clearly defined remit and goals, matched by defined level(s) of service which are achievable within budget, and by clear policies and procedures; there should be planning (short-, medium- and longer-term). (Recommendation 9.1)
3. Work needs to be sustained on the development of vocabularies and ontologies. These also need to be developed in the context of sustainability (a) to track changes as fields of research develop, and (b) to include features which permit translation of older terms. (Recommendation 4.1)
4. Resources in some instances (particularly community resources) need to be more aware of the needs of archiving and long-term preservation. We suggest that the sponsors engage the Digital Curation Centre in approaching this and take the matter up with the managers of those resources they support. (Recommendation 7.4)
5. Encourage the gathering of user input into tools development programmes. (Recommendation 5.1)
6. A code of good practice should be established for managing and sharing confidential data, which recognises both risks and benefits of sharing for all stakeholders. (Recommendation 8.3).

PART 1: BACKGROUND TO THE REPORT

1. Horizons and background to report

Digital technologies are transforming the life sciences, enabling scientists to collect and record vast quantities of data, analyse these, replicate them and distribute them to the scientific community through multiple channels, beyond the traditional peer-reviewed journal. In this way data, and tools and standards for working with the data, have extended the horizons and ambitions of the life sciences. This enables new orders of collaboration both within and between disciplines, based on the sharing of data and of tools to manipulate it and work at composite information level. This sharing is a major engine of discovery and success in the life sciences.

Purpose of the report

Such changes do not come without difficulties, not only for the scientists as they participate in this new science, but also for the technologists, computer and information scientists, and for the funders of these newly collaborating scientists.

The sponsors of this report (see Box 1) are concerned that maximum and optimum use be made of expensively generated and expensively maintained data to support excellent and ethical science in primary and secondary research.

They have therefore commissioned this study to examine data sharing, looking at user needs and best practice, identifying incentives, facilitators and obstacles, and, in the light of the findings, considering models which support effective, ethical data sharing.

Box 1: The study sponsors



The Biological and Biotechnology and Biological Sciences Research Council (BBSRC)

Department of Trade and Industry (DTI)

Joint Information Systems Committee (JISC) for Support for Research (JCSR)

Medical Research Council (MRC)

Natural Environment Research Council (NERC)

The Wellcome Trust

The scope of the study spanned the whole spectrum of the life sciences, from genome to ecosystems, and all digital data types and methods, from bench science to field surveys, including data-based science. Comparisons were made with other knowledge domains and with industry. PI-led case studies were to be covered as well as community resources, and the international context examined.

Premises underlying the report

The underlying premise of this study is that data sharing is important for research. The study was not required to establish the case for this *ex ante*. A number of recent studies have presented the validity of the premise [89, 90].

2. Method

The area to be covered by this report is vast. Rather than attempt a systematic study of the many component domains and resource types - a huge undertaking – the terms of reference of this study set a case-study approach on a small number of selected data resources, supported by interviews with key informants (see acknowledgements list) and desk research. A formal survey from which firm statistical inferences could be drawn was not sought.

The terms of reference pointed to specific domains, data and resource types (see Box 2 below), examining in particular:

- User needs in data sharing
- Incentives and barriers to data sharing
- The creation and implementation of data-sharing tools and standards
- Models of effective practice.

Box 2: Domains and data types covered by the study

<p>Domains:</p> <ul style="list-style-type: none"> ■ Human, animal, plant sciences, primarily focusing on: <ul style="list-style-type: none"> - Post-genomic research - Population research (health, disease, biodiversity) ■ In addition, selected exemplars from the environmental domain and comparators from outside the life sciences. 	<p>Data types:</p> <ul style="list-style-type: none"> ■ Genome sequence ■ Micro arrays ■ Biological structures ■ Laboratory and clinical data (including imaging) ■ Field and population
---	--

The case studies were selected to cover the domains and data types shown in Box 2, including both community resources and principal investigator-led activities. The span covered projects of different

ages and a range of points along the data-sharing/data management flow, from different perspectives, and illustrating different organizational models.

The report looked at the following ten case studies, in alphabetical order:

1. **AHDS:** Arts and Humanities Data Service – non-life-sciences comparator case study
2. **BADC:** The British Atmospheric Data Centre, including consideration of the NERC Data Grid (NDG)
3. **BRIDGES:** Biomedical Research Informatics Delivered by Grid-Enabled Services
4. **CLEF:** Clinical e-Science Framework
5. **CDA:** CDA Limited (Common Data Access) – commercial comparator case study
6. **Ensembl:** annotated genome sequence resource
7. **GIMS:** Genome Information Management System
8. **Malaria:** *Plasmodium falciparum* resource
9. **NASC:** The Nottingham Arabidopsis Stock Centre
10. **PSI:** The Proteomics Standards Initiative

BRIDGES was used as pilot case study. The study also looked at the following projects as mini case studies: OpenEHR, eBank, GBIF, neuro-imaging.

Further information on the case studies is provided in Appendix 3, including the rationale for inclusion and case study reports.

Findings are discussed in Part 2, and conclusions and recommendations in Part 3. Brief definitions of key concepts in the report are provided in Chapter 3. A bibliography is provided in Appendix 1, ordered alphabetically (by first author name). Technical and scientific terms are defined in the glossary in Appendix 2. This appendix also contains a list of acronyms and their meanings.

The findings of the study are reported here in the form of examples of good practice from the case studies, interviews and desk work, with commentary. In such vast territory, the findings do not attempt to provide statistical rigour.

A note on the examples and comments

Examples quoted of good practice are selective, they are **not** given for every case study in each instance; if a case study is not quoted as an example of good practice, that does not in any way imply deficiency.

All comments by informants have been anonymised.

The study consortium

This study was conducted by a consortium consisting of the Digital Archiving Consultancy Limited (DAC), the Bioinformatics Research Centre at the University of Glasgow (BRC), and the National e-Science Centre at Glasgow (NeSC). Key informant interviews and case study visits were conducted

by Alison Macdonald, Philip Lord (DAC), Dr Denise Ecklund (NeSC), Dr Richard Sinnott (BRC, NeSC) and Andy Jones (BRC). Dr Martin Westhead (NeSC) conducted key informant interviews and research in the USA as well as UK. Professor David Gilbert of the BRC provided advice and guidance. All contributed to desk research and content; final production was done by the DAC. The consortium was led by the DAC.

Conventions

Bibliographic references are referred to in the text by number in square brackets (thus: [1]). The bibliography is listed in Appendix 1.

Quotations are provided in italics, thus:

“This is a quotation” – Anon.

Quotes from key informants are anonymized.

Recommendations are numbered sequentially, based on chapter number, and are shown thus:

Recommendation 1.2: **Recommendation text.**

Examples of good practice are highlighted by use of a sans-serif font, thus:

In this case good practice . . .

In the main body of the report (chapters 3 to 15), chapters begin with a box listing key points and end with a short list showing related sections.

3. Key concepts

- Three important naming conventions are used through the report:
 Producers: Individuals or organisations which generate data for sharing.
 Data resources: Organisations who are intermediaries between producers and consumers, storing and making data available. A data producer can also be a data resource.
 Consumers: Individuals or organisations who receive data.
- It is not just data that needs to be shared but a variety of other digital objects such as metadata, software, processes, method and process descriptions, vocabularies.

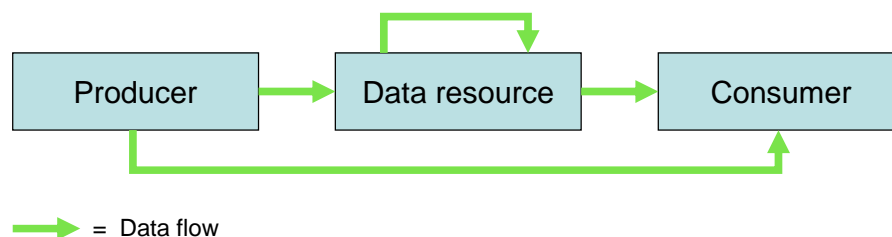
In this chapter we provide brief working definitions of the key concepts discussed in the body of the report.

Sharing and standards

Sharing: In this study we take data sharing to mean the **re-use** of data, in whatever way, and wherever it is stored. (Re-) use includes actions on data objects such as consultation, visualisation, (re-) analysis, transformation, annotation, enhancement, etc.

Sharing can be classified by time: **synchronous** sharing is when re-use is more or less contemporaneous with data creation; **consecutive** sharing, when data are passed from one person or body to another (this form of data sharing is common in the pharmaceuticals and oil industries); or **asynchronous** when shared after longer intervals, perhaps amounting to decades or centuries (as when archived). Of course, digital media also allows sharing in parallel by many people at any one time. Sharing in a clinical care context involves all three types of sharing.

Sharing involves a **producer**¹ who is the source of what is to be shared (often its creator), and a **consumer** (sometimes called a customer, user or recipient). In some cases one or more **data resources** which store and/or make data available may lie between the producer and the consumer.



Sharing can happen where consumers are sent information, by a **push** from the producer or data resource, or the consumer can seek the information to be shared, and **pull** it to him or herself.

One further distinction is that sometimes resources are shared **in situ**, where data are exploited on-line while staying in the resource; on the other hand **remote** sharing takes a copy, or extract, of data from a resource and uses it locally. Both modes can be mixed in any one sharing instance.

Standards: Standards relevant to this study are those which describe norms which should be adopted to facilitate interchange and inter-working of information, processes, objects and software.

¹ These terms are used in the Open Archival Information Systems standard referred in Section 7. See also the DCC website (<http://www.dcc.ac.uk/>); the DCC also use the term “Data Re-User” for consumers.

They are discussed further in chapter 6. Standards for levels of quality, management, behaviours are less directly relevant but underlie the effectiveness of data sharing.

Prerequisites: Sharing requires some fundamentals to be in place: to pull information it must be **discoverable** through mechanisms such as indexes, catalogues, search engines, portals etc. and these tools need to be discoverable and available to the consumer. In case of difficulty, it is preferable that players in the sharing have access to **support** facilities.

For all this to work, a reliable computing and network infrastructure has to be in place of networks, access and authorisation systems, as well as computers.

What is shared

What is shared is data itself, annotations to the data, processes, methods, processes. In general data alone is not sufficient for its re-use: To be able to use somebody else's data with ease and confidence, you need to know something about the data and how to process and interpret it – for example, information about the experiment and its context, technical information about the format of the data, what software to use. This information about data is called **metadata**.

Data: We focus on data in digital form, whether raw data newly gathered (by whatever method, including simulation and modelling), or data derived from that by processes of consolidation, selection, calculation, and statistical analysis, etc.

Metadata: This is usually defined as data which describes or qualifies other data. Here it includes descriptions (of all kinds), annotation, documentation and commentary.

Tools: As a rule in this context we mean software which in some way facilitates the process of sharing data: it thus includes software for data discovery, visualisation, processing in the context of re-use, storage, archiving and preservation, and the management of sharing processes.

Methods: Some data may make no sense when shared unless the methods and workflows used to process and view it are also shared; usually these will be delivered to the recipient in the sharing process either as metadata or encoded/implied in software tools.

How it is shared

In this report “**large-scale sharing**” means that one or more of the following hold:

- volumes are large (gigabytes and beyond)
- the number files or records are very large (e.g. in millions)
- the number of transactions which take place is large
- the number and/or diversity of participants is large.

We distinguish different kinds of data resource. A **repository** is a facility to manage the storage of data, possibly providing an index to its contents. A **community resource** is an actively used data resource which stores, maintains, curates and makes available data-sharing resources as defined above

for a specific community of users (as for example the NASC does for the academic community working on plants). There is a spectrum of provision between these extremes.

Other characteristics distinguish data resources: some offer datasets in the form of discrete files for downloading (“**data farms**”); others offer **integrated** databases where one can drill down to specific data items (such as a short base pair sequence in a genome).

Curation: The activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and with other published materials [69]². The term was introduced by Dr. John Taylor, when head of the UK’s Research Councils, who defined it as “maintaining the intellectual content of data”. The concept includes archiving data and preserving it over the long term.

Some subject domains use the term curation in different ways. Of significance here is the special meaning attached to this activity in the field of genomics, where it is used mainly to indicate enrichment of sequence data information or other commentary. UniPROT and the Malaria *Plasmodium falciparum* database at the Sanger Institute are examples of a “curated” database.

Players and roles

Data sharing takes place within multiple contexts, involving many players who assume a variety of roles. Further aspects are discussed in Appendix 4. The major players are merely listed here:

- Funders of research
 - Umbrella funders (i.e. the funders of funders)
 - Communities at large – domain-based. E.g. the *Arabidopsis thaliana* community, which is part of a wider plant science community. Communities can be overlapping.
 - Standards bodies, formal (e.g. ISO) and informal (e.g. MIAME); national and international.
 - Data resources: those who manage data, applying varying levels of curation (repositories through to extensive curation). Allied to this community are libraries and archives, and those community resources that simply provide information on other data resources.
 - Publishers, commercial and not-for profit, including open-source publishers
 - Intermediaries who provide direction and/or support (such as JISC, the e-Science Programme and the Digital Curation Centre), or which provide infrastructural research (such as the National e-Science Institute)
 - Those who are performers in the information-sharing chain: producers, consumers (severally as individuals, teams or research organisations); data resources as defined above.
 - Research host institutions, such as the universities.
 - Others, such as learned societies, patent offices.
- All these players now operate over an international canvas of considerable complexity, as illustrated by the diagram in the NASC case study (Appendix 3.9).

² This reference discusses and defines digital curation further. See also similar definitions on the Digital Curation Centre’s website, <http://www.dcc.ac.uk/about/what/>.

PART 2: FINDINGS

Each section which follows begins with a summary, and a “route map” to related sections is provided at the end of each.

Good practice: The criteria for citing examples of good practice are derived from the case studies, informants’ views and desk research. The criteria are that the examples do one or more of the following:

- Increase the ease with which data can be used by consumers
- Reduce the costs and complexities of the operations performed by data producers, resources and/or users
- Help producers with their tasks of providing data and metadata in a form which will increase their value for sharing and lighten the burdens on data resources and consumers
- Promote longevity of data in a usable form
- Facilitate funders’ wish to make best use of the data assets whose creation they support
- Promote ethical and legal reuse of data.

Guide to subjects discussed

Subject	Sections	Subject	Sections	Subject	Sections
Archiving	7	Funding, costs	9, 13, 15	Preservation	7
Careers	10	Guidance	11	Processes	4
Confidentiality	8	Incentives	10	Quality (of data)	13
Consent	8	Industry	13	Regulation	8
Culture	10	Infrastructure	4, 5	Security	5
Curation	7	IP Rights	12	Selection	7
Data management	5, 7	Legislation	12	Service delivery	9
Databases	4	Location (resources)	9	Software quality	5
Development (tools)	5	Metadata	4, 7, 10	Standards	4, 6
Discovery (of data)	4, 5	Model – data flow	15	Tools	5
Digitisation	7	Open source	5, 12	Training	11
Ease of use	5	Organisation	9	Unresolved questions	15
Ethics	8	Planning (data management)	7	Vocabularies	4

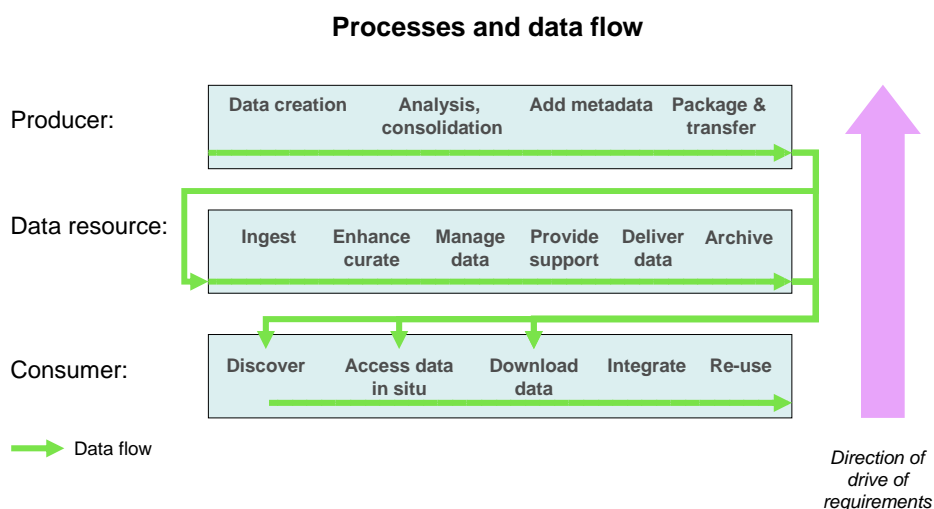
4. Technical issues

- Technical difficulties in re-using data are substantially eased by good data management, notably timely collection and provision of data about the data:-
- Data generators need to provide adequate, accurate metadata to their digital output at time of creation for efficient, cost-effective quality data sharing downstream
- Lack of interoperability, semantic ambiguity are major obstacles to data sharing; these are addressed by standardisation, tools and efficient administrative and data processes

This section presents some of the major technical obstacles to data sharing encountered in case studies, interviews and desk research, with examples of good practice. The technical components of these issues fall broadly into two groups, computing interoperability and semantics.

However, several obstacles at first sight apparently technical in nature were shown to be wholly or partially issues of standards (section 6), administrative process (see section 7 and 9), or behaviours (see section 10).

The diagram below shows the major processes performed by the three groups who play an active role in sharing: producers, data resources and consumers (as defined in section 3 above). It also shows the flow of data from creation to re-use. However, in terms of the needs for efficacious data sharing, requirements tend to flow in the other direction: from consumers to data resources and producers (in commercial terms, “giving the customers what they want”).



We present this section following the flow of **requirements**, from consumer to producer, noting the points in the data flow where the major issues lie.

4.1 Discoverability - semantic issues

Discovery refers to the process of finding information of value by consumers from data resources. Discovery can be a major hurdle for consumers to cross. It requires software tools, such as search engines. While Google is universally used, different data resources may need specialist tools to serve

different communities' specific needs. Discovery tools include portals, vocabularies, ontologies, and data-specific software such as those for probing genetic and proteomic databases and then displaying (“visualising”) or extracting information.

Discovery also relies on descriptive metadata and the absence of confusion caused by semantic ambiguity. The latter is addressed in part by controlled vocabularies and ontology tools. Each of the case studies and many interviewees provided examples for these: from the corporate sector there was a simplified example from documentation used in industry – the use of “tap” in the UK and “faucet” in the USA. Without a tool which was semantically aware, a search for taps in a ship design plan prepared by an American for a multi-national engineering company would indicate no taps, so none would be ordered by the builder.

The need for agreed vocabularies was raised many times by academic and commercial groups. For example, lack of an agreed vocabulary for atmospheric chemistry datasets from NASA at the BADC means they cannot be used with confidence and consequently they are not made available to consumers.

The reverse problem also exists, in the corporate sector and clinical field, with parallel or competing vocabularies developed by different bodies.

A further issue is that changes in technologies and simple usage cause changes in vocabularies and meanings, raising the issue of currency of vocabularies and their maintenance.

Recommendation 4.1 Work needs to be sustained on the development of vocabularies and ontologies. These also need to be developed in the context of sustainability to (a) track changes as fields of research develop, and (b) include features which permit translation of older terms to newer.

Good practice

- Discoverability: NERC DataGrid (case study, Appendix 3.2) specifically addresses all these potential obstacles (note: part of the problem it addresses is rooted in poor practice at data management level).
- The neuroinformatics community has a portal, the Neuroscience Database Gateway, (<http://big.sfn.org/NDG/site/default.asp>) run by the Society for Neuroscience, to provide access to the burgeoning number of databases available to the community. Its first version numbered over 70 databases and related resources. It provides a structured access method, supporting navigation around databases, disciplines, levels of organization and laboratories, with tools for users to conduct sophisticated searches. [3]
- CO-ODE, (www.co-ode.org/) developed by the University of Manchester Medical Informatics Unit, is a practical tool which helps users build ontologies (“an ontology defines the terms used to describe and represent an area of knowledge. Ontologies are used by people, databases, and applications that need to share domain information”, definition source W3C).

4.2 Databases

Data in databases can be accessed and used by consumers in several ways:

- Through data provider tools such as web front-end forms, through which queries can be made to drill into databases in situ – this is arguably the norm right now for most scientists
- Downloading whole databases or parts of them and using them locally, individually, in integration with other data, or for comparison – this requires local knowledge and expertise by local staff to ensure consistency with remote (and evolving) datasets
- Accessing collections of remote databases directly by using targeted applications and associated middleware which overcome many of the issues in remoteness and heterogeneity of resources and associated data models; this is now feasible, at least in principle, through Grid middleware
- Combinations of these approaches are also possible, and are often necessary since the programmatic access needed to use Grid³ middleware to facilitate federated access to databases, is often not provided although the datasets themselves are made available (typically via compressed flat files). Here again though, the schemas associated with this raw data are often not provided.

Once again, interoperability at computing and semantic level is key to easy sharing of database data. Standardisation of methods, syntax and semantics (within fields and in labelling) makes searching and comparison across multiple databases more accurate, richer and easier. Currently such standardisation does not *fully* exist in any given life science domain, although numerous communities are making efforts to address this (Gene Ontology, MAGE-ML). As a result it is often not possible to find a common table name or row over which a federated query may subsequently be joined.

In addition, not only are many of these databases constantly being updated, as science itself is advancing and new insights gained the databases and associated data schemas themselves are evolving. Applications using such federated and evolutionary data models will break, with exceptions being raised when incompatibly formed queries are issued. Whilst science and the associated data sets and data models will continue to expand and should not be prohibited from doing so, users and application developers should be informed by data providers when such changes are to occur. This is predominantly not the case at present.

More technical aspects relating to databases are set out in Appendix 5.3.

Recommendation 4.2: Guidance and incentives should be given by funders to encourage those creating databases to document database schemas, naming conventions clearly.

Good practice

- ☑ The EBI, the EMBL outpost, is home to over a hundred databases connected together. From the outset the institution has been collaborating on data exchange between databases with other key database resources, in the USA and Japan. As a result, a query run on the EBI web site can search seamlessly across diverse databases at a range of different institutions.
- ☑ Users of the SRS search technology (available on the EBI web site and commercially) make details of their database structure and data parsers (see glossary) available to the system, so its interoperability is constantly extending.

³ Please see the glossary in Appendix 2 for brief definitions of technical terms, such as “Grid”, SRS, etc.

- ☑ Malaria *Plasmodium falciparum* is one of the genomes available in the GeneDB database family available at the Wellcome Trust Sanger Institute's Pathogen Sequencing Unit. Through a single portal users can access, visualize, search and download data, supported by a database architecture which allows integration of diverse datasets and by structured vocabularies (using the Gene Ontology, see malaria case study), facilitating querying and comparison between species. The GeneDB team has been working on integrating its database with the GUS schema (the Genomics Unified Schema, developed by the University of Pennsylvania, see malaria case study), a platform used by several of the major biological databases and available to the community. This is an example of the co-ordination and conscious use of common tools and structures to enhance user access to data resources.
- ☑ The BRIDGES project has developed services that detect when remote data sources have changed and automatically download and parse the data files for input into the local data warehouse.

4.3 Re-use of data

Re-use covers a vast range of activities: calculation, reanalysis (such as statistical analysis, including meta-analyses on aggregated datasets), visualisation, modelling. Many of these will be domain-specific, thus, as one example, image analysis to extract features – say tumour identification in a radiograph.

Standards facilitate data re-use. They make data sharing easier, saving overheads and losses of time in data loading, conversion, getting systems to work properly with data received, and with interpretation. Standards *per se* are discussed further in section 6. Standard data formats either come from the producer or are provided by the data resource following conversion after ingest.

4.4 Delivery and ingest – the infrastructure of sharing

“Ingest⁴” to data resources from producers and delivery to consumers are facilitated by networks and computers with adequate capacity for transfer and processing. This is a matter of infrastructure. After the publication in 2004 of the government's Science and Innovation Investment Framework 2004-2014, the OST has started to explore equipping the UK with a first-class “e-Infrastructure” for research, recently appointing a steering committee to define requirements and steer progress⁵. This e-Infrastructure, within which data sharing in the UK will operate, defines an infrastructure which is wider than just networks and computer hardware. It also includes: information and data itself (including vocabularies and ontologies); tools (software for data creation and collection, for data discovery, middleware for collaborative, inter-disciplinary working, and for research management); authentication, authorisation and accounting mechanisms (“AAA”); support services; preservation and archiving services.

All of these are essential ingredients in the whole infrastructure. Many of these elements are referred to in this report. Data-sharing tools at the level of infrastructure are discussed further in section 7. All the elements are dependent on the establishment of standards, most of which need to be set on an international scale. These components of the e-infrastructure will be variously provided, on local, national and international scales, as appropriate.

⁴ Ingest, an OAIS standard term, is the whole process of sending data from a producer to a data resource and preparing it for subsequent discovery.

⁵ The steering group has representation from OST/DTI, JISC, RCUK, HEFCE, CCLRC, EPSRC, NERC, the British Library, The National Archives, the Research Information Network, and industry.

The definition and establishment of this e-Infrastructure is not discussed in detail here, being beyond the study's scope. Clearly if infrastructure is deficient – say in network capacity or security of access to resources – then it has the potential to hinder data sharing.

4.5 Metadata

The term metadata is used variably by different research communities. Box 3 provides a short overview, but it may be summarized as providing the necessary evidential and context information to enable data to be correctly interpreted and used in a wide range of situations, and over time.

The origin of many of the difficulties described above lies in inadequate metadata. As already stressed, this was a recurring theme from case studies and informants (and backed by desk research) that consumers need sufficient supporting information to enable them to load and use data, information which provides them with indications of its source and trustworthiness, and information which ensures they can interpret it correctly. Lack of interoperability between different metadata schemas, structures and labels can also compromise sharing.

The unanimous consensus of all those professionally involved in data curation and sharing is that descriptive and contextual metadata is best provided by the producer as part of the data-generation process. The BADC noted that creation of metadata is a “people problem” because:

- It cannot be fully automated
- Is a recurring human labour-intensive activity that must be carried out (and thus paid for) each time new data are created.

Box 3: Metadata types

Suitable metadata has a number of components:

Descriptive and contextual metadata: describing the technical characteristics of the data and its formats; data structures; the context (such as experimental set up, parameters) and provenance (who created it, where, when, and for what object). Some simple examples: File formats used (such as JPEG); the date and time of an experiment, the laboratory name, the producer's name, instrument calibration parameters, the design goal of the experiment, . . . etc.

Semantic information: the meanings attached to labels, data elements, and the vocabularies used. Examples are names for genes, field names in a statistical database, protein names.

Identity of special software tools or processes needed to use or interpret the data.

In some cases workflow descriptions may be relevant to future understanding and use.

Information which facilitates subsequent discovery of the data by consumers, such as keywords.

Data resources that enhance the data they receive generally add further metadata to that supplied by the producer (such as data received, status ...etc.)

Archiving and preservation need further metadata, such as specifying how long to keep data and the technical underpinnings of the data so as to enable preservation to take place. Currently It is not always clear who provides these (q.v. section 7.7).

Behind these problems lie questions of process, co-ordination, the availability of resources (time, training, guidance) and human factors such as motivation and incentives, discussed in sections 9 and 10 below.

Recommendation 4.3: Developing better systems for collecting metadata, either manually or automatically is one of the most urgent issues needing attention and research. Resources should be directed in a concerted fashion to this problem.

Recommendation 4.4: There is a need to develop a degree of standardisation in the way metadata is represented and structured to enable interoperable processing and sharing.

Good practice

- Resources are trying to address improvement in the quality of metadata provided by data producers by attempting to get it assigned early in data's life cycle: thus BADC is encouraging this, for example. .
- A good example of developing clear categorisations and structures for metadata is the NERC DataGrid. This will categorise metadata into four groups:
 - o A [Archive]: Format and usage metadata. This supports physical access to archived data
 - o B [Browse]: The superset of discovery, usage data, and contextual metadata. It supports browsing and selecting of datasets
 - o C [Comment]: Annotations, documentation and other supporting material, including comments from the data creator
 - o D [Discovery]: Metadata used to find datasets. It will support data discovery (In the NGD it will be conformant with ISO 19115, which defines the schema required for describing geographic information and services)

Considerable work on metadata has been done by the Nerc DataGrid team, and the BADC highlighted the metadata work of the BODC. More generally, the BODC site is another example of helpful, clearly written materials to support the user.

- The Intra-governmental Group on Geographic Information (IGGI) has issued a useful booklet for researchers on the principles of good metadata management, with wider applicability beyond the geographic sciences.

Related sections

Standards	Section 6
Tools	Section 5
Process and resource management issues.	Sections 7, 9
Data planning	Section 7
Archiving and preservation	Section 7
Behavioural issues	Section 10

5. Tools

- Tools of value to the scientific community should be properly developed, tested and maintained.
- User input to tool design and ease of use are key to use of the tool.
- Small demonstrator projects are of enormous value and minimize risk.
- Interoperability of tools increases their value to users.

5.1 Tools underpin and add value

Digital tools enable, facilitate, incentivize and enhance data sharing. They do so at all levels - infrastructure, administration, data management and science.

At **infrastructure** level tools facilitate data sharing, for example by providing security and controlling access. They are critical to data sharing in areas which work with confidential material, such as clinical research, or where rights must be protected. The need to address “AAA” (authentication, authorisation and accounting/auditing) is paramount to engage the medical community in sharing of data and the access to and usage of resources (computers, databases) which host such information.

With project members from several organisations, the BRIDGES data warehouse needed to provide differentiated access control to reflect different privileges for different data for different users. As several key informants stressed, access tools need to be able to differentiate between roles and functions rather than between individuals: the same person might be a university researcher in the morning, but a director of a biomedical start-up in the afternoon, with very different entitlements. The PERMIS role-based access control software has been used for this purpose. A key factor in exploring these technologies within BRIDGES was removing as much responsibility away from the biomedical end users as possible. The process of acquiring X.509 digital certificates and their subsequent usage and management can be involved and a turn-off for non-computer scientists. BRIDGES has recognised this and through portal-based solutions using user name and passwords, provided solutions whereby scientists can for example securely share data and gain access to resources such as the National Grid Service without explicitly having their own digital certificates.

While these tools provide, ideally, seamless access that corresponds to entitlement, privileges need to be agreed and maintained. There is an ongoing administrative element to their provision. Fine-grained security policies defining which roles can access which data sets can be implemented, but there are certain drawbacks. Firstly the data sets are likely to be evolving, or more problematically the schemas are changing. This will make the security policies redundant. Secondly, the details of these security policies can rapidly become non-maintainable. Defining security policies at detailed level has clear scalability issues.

At the **discovery** and **data management** level, tools are essential. In the life sciences domain special tools are needed to search within and across specialist data types such as gene sequences. In Europe, the EBI provides a toolkit of such tools; users can also use an EBI service which runs searches for them applying powerful computational processes, applying power which may not be available to the user.

As we have seen in Section 4, data discovery and data use depend on the availability of metadata. Tools which facilitate the collection and management of this data mean metadata is provided in the first place, and also mean that better metadata is provided in more efficient manner and format. They also increase the amount of time available for actual research and increase the quality of the data, at research project and repository level.

As one key informant stressed, tools also provide the incentive to use data, enabling the researcher to leverage the data for his research:

“Data sharing [...] brings previously disparate pieces of data together for further analysis. [...] Without new integrated analysis tools to take advantage of the integrated data, scientists may not receive significant gains in their research.”

Developing these value-added tools, however, often requires specialist IT expertise which the researcher is unlikely to have.

Thus, in the biological sciences (for example) the bioinformatician is an intermediary. Proximity and good communication between informatician/ computer scientist and scientist are very important. Input from the researcher is key to the utility of the tool. However, some key informants said that it is sometimes difficult to get the researcher to sit down with informaticians, to clarify points relating to tool design. One set of key informants ran a training course on an existing tool:

“this provided first-hand feedback on new features which users wanted most, and the tool team were able to add these features in a very short time.”

This resulted in much higher use of the tool and also a shift in attitude on the part of researchers to communication with tool developers.

Recommendation 5.1: Encourage the gathering user input into tools development programmes and projects.

5.2 Quality of tools

Tools are often developed specifically for a particular research project and then made available as open-source tools. More often than not, these tools are prototypical in nature. Several informants pointed to a lack of “robustness and resilience” in these tools – they tended to break down and did not cope well with heavy loads. As one key informant explained,

“This is because they tend to be hastily developed by a single developer (a) with not enough time, and (b) with no experience of developing robust software, with documentation, testing and validation of the application. Industry is not interested in purchasing the technology ... in part because its design and development have not been rigorous.”

The documentation about the tool is often not attached to the tool itself, which makes it hard to find and difficult to use.

Similarly, it is not enough to pay for prototype or research-level development of a tool, make the tool open-source and expect the rest to happen. A large proportion of the cost of software application is in its ongoing development and support. If tools are providing value to a scientific community, they need to be properly tested and maintained. The UK e-Science community has recognised this and

funded the Open Middleware Infrastructure Institute (OMII) whose remit is to make leading Grid middleware more robust through re-engineering.

Another key informant drew these strands together:

“Sharing tools is also important. Tool innovation comes from the university, but the resulting technologies are not robust and industrial strength. If the bio community is to share tools, then those tools must be robust, reliable, well documented, etc. To this end funding is required for someone to “drag the technology through” beyond the prototype stage and to maintain the tool.”

Tools such as those developed within a specialist unit or those provided by the bodies such as the EBI or NERC’s Environmental Genomics Thematic Programme Data Centre (EGTDC) have the development expertise to make the tool robust and resilient. They maintain a balance between enough testing and maintenance for the tool to be of practical value, and not over-investment of resource given the pace of change in IT technology and the evolution of research.

Recommendation 5.2: When funding the development of tools, set standards for the quality of the software developed, appropriate to the need and the level of use envisaged.

Good practice

- Resources such as the EBI and NERC’s Environmental Genomics Thematic Data Centre (EGTDC) combine scientific and informatics expertise, and develop tools for their communities; the EGTDC was one of several examples where they extend existing, tried and tested software specifically for the environmental genomics community.

5.3 The importance of demonstrator projects

Many key community resources were born from demonstrator projects and informal ideas. Ensembl is one example. Another is the SkyServer, which makes the entire public database of the Sloan Digital Sky Survey (over 80 million stars, galaxies, quasars) available to the public free of charge, through web pages. The site includes tools to view and download the data, and links to sample applications. The SkyServer Survey grew out of a six-week student project, which quickly demonstrated the benefits of the project.

The value of demonstrator/prototype tools projects cannot be overemphasised. Many scientists will be able to produce a long laundry list of requirements for their data sharing, but there are few that can get the priorities right before they have seen a working demonstration. Producing limited functionality, working code is immensely important in bootstrapping the requirements processes. Small projects also help to minimise risk. If the Sky Server project had failed to produce anything useful, very little was lost.

5.4 The importance of key technologies

Part of what made it possible for SkyServer to produce its product in such a short time was the availability of a generic platform on which to develop. Many of the underlying technological issues

in data sharing are common, regardless of the application area. The construction of infrastructure software to support these generic areas represents an important cost saving because a single technology and support infrastructure can be used by multiple projects rather than each project having to start from scratch.

5.5 Ease of use

Ensembl is an example of a suite tools designed to be easy to use. The tools cater for different levels of user expertise, from layman or beginner to experienced user.

Behind an apparently simple façade lie highly complex workings. This point was repeated by more than one key informant:

“Tools must be simple for the biologist, but at the same time these tools must model a complex process”.

Tools modelling a complex process which are also easy to use are usually hard to put together. To do this calls for a combination of skills and knowledge: the developer needs to be a computer scientist, probably also an information scientist, and a scientist with an understanding of the relevant domain. The success of the Ensembl effort to achieve usability is reflected in the enormous level of use – in 2004 the web site served in excess of 500,000 web pages per week.

Ease of use is a key factor in level of use. We came across many examples where, when confronted with additional labour or an unfamiliar activity, there was resistance to tools developed for them. A typical statement:

“ .. These tools were not popular because the biologist was required to run a few PERL scripts from the command line.”

Training, of course, plays a key role, which is discussed in section 11 below.

Ease of use is often equated with presentation in web-based form. The conceptual model which underlies the web-based façade, however, can be quite different from its appearance as mediated by the typical web browser paradigm. This can be a source of confusion. The Ensembl team has occasionally found that some users express their queries in terms of the web interface, only to find that if the question were framed in terms of the underlying scientific question, their query is answered.

For inclusion of a database in its annual listing, one of the major journals requires that the database can be used via the web. However, using a web interface can inhibit the power of the tool. To overcome this requires substantial engineering. Another key informant noted that a large volume of life sciences data is complex in form, often images, requiring specialist analysis tools which require a level of computing power which is difficult to deliver in a web format. Here again, the importance for more general computing training is evident.

5.6 The importance of integration

Tools need to work together. This is probably an obvious point in the area of data sharing but it is so important that it deserves special attention. The value of each individual tool can be significantly multiplied if it can be integrated into a consistent toolset so that users can leverage their

understanding, parameters, data, models etc. across the whole set. This point speaks to the importance of using standards but also encourages developers to work together and use common frameworks and technologies.

The point was reiterated by a key informant in the context of support tools:

“Without a common set of supported tools bioinformaticians working in small research project must reinvent these applications. This is wasted effort and won’t be done. Hence, biologists will be less able to analyse their data and scientific progress is hampered.”

5.7 Open source and multi-platform issues

Many informants stressed the critical importance of open-source licences in development of tools for scientific data sharing. Advantages of open-source licences include:

- communities of users can share the support and development costs
- specific requirements/modifications can be added or made to the software on the timescale of the project (at the project's expense)

However, software which is open source or free may not always remain so. A change to a charging model for software could have a substantial impact on budget, either directly by accepting the new model, or indirectly by switching to alternative software. Projects and resources should endeavour to ensure that software and tools that can run on multiple platforms, to minimize the impact of change in charging status.

Good practice

- ☑ The EBI provides a toolkit of tools, in step with evolving research needs, and incorporating feedback from the international community.
- ☑ NERC's Environmental Genomics programme provides software, support, consultation and training as well.
- ☑ maxdLoad2 is a standards-compliant database development which supports the export of microarray data in the MAGE-ML format, for submission to the Array Express database, to comply with journal requirements. A PI-led project developed by Manchester University Bioinformatics department, the maxd web resource provides the software, with clear information on installation and use. The tool, with support, is provided by NERC's EGTDC to its community.
- ☑ Tools re-use: Ensembl software has been used for many other projects, including the NASC Arabidopsis resource and the CADRE Aspergillus repository. The NASC team closely worked with Ensembl in the development of the Arabidopsis Ensembl tool. One advantage of re-use is that the Ensembl software is tried and tested: it has been through previous versions, had feedback from users, errors eradicated.
- ☑ NERC DataGrid: part of the project addresses the need for documentation to accompany software tools. The BADC provides a suite of common data analysis tools and models for its users, and works with the research community to determine what applications or tools are needed. In the BADC example, geo-analysis tools are developed either by BADC or by other researchers who agree to make their tools available. They offer web-based access to tools and data.
- ☑ The BRIDGES project has developed a Grid-based infrastructure supporting both a data Grid and a compute Grid for biomedical researchers. Many of the core challenges in data sharing and security through Grid technologies (both commercial and open source) have been investigated within BRIDGES. It is expected that many of the BRIDGES solutions will be further supported in other research projects such as the Scottish

Bioinformatics Research Network and further research proposals. It is also the case that BRIDGES are also driving much of the bioinformatics work on the UK e-Science National Grid Service.

Related sections

Metadata - needs and provision

Section 4

Data discovery and re-use

Section 4

Training

Section 11

6. Standards

- Standards are fundamental to data sharing, at all levels – administrative, data and science
- While community participation in standards development is essential to their uptake, development must be rigorously controlled to avoid bloat and over-complexity which lead to rejection of standards
- Information/data standards need to be flexible to serve a wide user base and adjust to rapid science and technology change

6.1 Why have standards?

Again and again, the obstacles to data sharing encountered had been or would be removed or substantially lightened by standards. For example:

“The need for the development of standards for clinical trial laboratory data is painfully apparent to those involved in the multiple laboratory data transfers that occur during a clinical trial. Some central clinical laboratories currently support up to 1200 different data formats. Every sponsor has developed a nearly proprietary data standard, and often has multiple standards across different therapeutic areas or business units. Each data standard requires development, validation, testing, support, training, and maintenance. For each player in the laboratory chain ... effort is expended in exporting data from one system, importing it into another system, verifying the transfer, correcting problems, retransfer of data, and appropriate quality control.” [CDISC, The Clinical Data Interchange Standards Consortium]

In summary, working in the absence of standards means increased use of resources, substantially impaired productivity, and ultimately increased costs. By introducing standards, CDISC is aiming to achieve multiple benefits – increased efficiency, savings in time and cost, and also streamlining the approval time for new products by the drug regulatory agencies.

*“Effective management of engineering information is hampered by the different computer systems, application programs and data formats used to create and store it. **These incompatibilities affect quantity [...], quality [...], exchange and sharing.** [...] There are several ways of overcoming these problems. **The ideal way is to use a standard way of representing engineering information in a form that is computer sensible (so that it can be exchanged and shared electronically), neutral (independent of commercial hardware or software), and international (rather than company, vendor, .. national).** With this “common language” you can exchange and share high-quality information between organizations and across the life cycle.” (Chris Angus, BT. Our emphases.)*

The second quotation mentions another form of cost, the decrease in quantity and quality of data which can be shared – an opportunity cost which cannot always be measured in financial terms.

The payback for the investment in standards is two-fold - the standard will promote wider and easier sharing of data at any one time; and it makes the data’s survival over time easier, as it will be easier to maintain and knowledge about the data (or tool) will be more likely to be available.

Box 4: Developing standards – STEP: detail to model

The context in which Chris Angus was writing was that of STEP (ISO 10303), the international standard developed for the exchange of technical product data, which it is estimated accounts for some two thirds of ISO activity. The evolution of the STEP standard is instructive.

During the 1970s several national protocols and standards were developed to address the issue of exchanging geometric data between different CAD (computer-aided-design) systems. By the early 1980s it was clear that a single more comprehensive standard was needed, international in scope and ideally covering the whole life cycle of use, and archive storage, of product and project data.

So in 1984 the International Standards Organisation (ISO) established TC 184/SC4 (Technical Committee 184, Sub-Committee 4) to undertake the work. Starting with three people, the committee grew over the next ten years to some 500 persons from about 20 nations. It took ten years before the first 12 parts of the new standard ISO 10303 had passed through the various committee processes and were ready for publication as full international standards (December 1994). The standards produced were complex and detailed.

In 1992 the DTI funded a three-year project, with matching industry funding from some 30 companies, the project continuing thereafter on a voluntary basis. From this work, the participants realized that they needed a generic data model. This has led to ISO 15926, a seven-part standard, incorporating a Data Model and Reference Data Library (RDL) - "a language for engineers that are processable by computers" – which is being adopted back into the STEP community.

6.2 Developing standards

Developing standards, however, is a difficult and slow business, and particularly difficult for technical data. Box 4 gives one example; the case study on the Proteomics Standards Initiative (Appendix 3.10) looks at standards development in more detail, charting process, scientific and technical difficulties, resources and strategy.

Initially, standards tend to emerge “bottom up”, from the community. Almost all encountered during this study were international and involved public and corporate sector. A corollary of their informal birth, however, is that development can take place without a formal support framework and without funding. Several informants confirmed that this puts a difficult task under substantial additional strain.

Email and the web are an enormous help in developing standards, providing rapid communication and ready access to pooled documentation. However, it is still the case that getting busy people from all around the world in one place at one time to work on and agree a standard is a slow business. But at the same time the context in which any one data standard is being developed nowadays tends to evolve at a rapid pace. The time scale highlighted by one key informant working in astronomy was the two-year evolution of web/XML standards, requiring a process that enables evolution and migration of astronomy standards which is commensurate with that external time scale. One key informant noted that there is a delicate balance between spending time developing new standards and spending time implementing them – if implementation is not fast enough, the standards will already have been rendered obsolete by new standards. A further point he made is that a key cost function is minimization of the number of lines of software code that need to be written by humans to keep up with changes in standards.

In almost all the standards reviewed, standardization required development of not one but several standards, in a tree of inter-related components.

Again, community participation in standards development is one of the keys to success and adoption by users. Some of the newer initiatives are able to make use of frameworks developed by other standards projects, as exemplified in the Proteomics Standards Initiative (see case study, Appendix 3.10).

- Recommendation 6.1: Generic guidance giving recommendations about management, structures and mechanisms in the standards development process, should be available and readily found by those embarking on or involved in standards development.
- Recommendation 6.2: Funding should be provided for those involved in standards development and maintenance.

- Funding for standards development in astronomy has been provided as a core part of research funding; in the UK this has been funded in large part by the AstroGrid collaboration which recognizes support for standards as part of its remit.
- The Virtual Observatory has adopted a leap-frog strategy for standards development and implementation, carrying out “throw-away” implementations of standards in advance of concluding standards discussions.

6.3 Pragmatism and rigour

Achieving consensus is not easy and takes time. Agreement needs to be reached across competing positions, whether proprietary products are affected or not. The standards development process needs to apply pragmatism and some rigour in standards definition:

“If you try to produce something which does what everybody wants, you find that when you want to use it, it doesn’t do what you want it to do.”

Several key informants gave examples where slack control of the standards definition process can end up with bloated standards:

“You end up giving everybody the right to add, but not to veto. [...] you need to gather requirements and prune ruthlessly”.

A key informant stressed the importance of not defining local rules as global rules [123].

Development requires firm chairmanship, a formal framework and good communications channels.

6.4 Flexibility

Standards relating to information are not like standards for an electrical plug design. All the information-level standards in this study found that, to be effective, they needed to be drawn at a high level. This allows extensibility and flexibility for the standard to work in multiple contexts. Standards for data, particularly those involved in processes, must bridge a wide range of different users, uses, and technologies. Lack of flexibility was identified by the CDISC LAB study [8] as one of the reasons why existing standards were not used in clinical trials:

“The structure of existing standards often does not accommodate the specific requirements of clinical trial laboratory data; [...] Some of the existing standards have limited flexibility that makes adaptation for clinical trial use cumbersome”.

A closer look at the characteristics targeted by the model being designed by the CDISC LAB group for a laboratory data standard for clinical trials shows that it is targeting two types of flexibility - for a breadth of users, and to keep pace with evolutions in laboratory testing in terms of model and technologies, over time.

6.5 Adoption: ease of use, role of journals

In the examples encountered, adoption of standards was higher when the standard was not too complex. Those standards with detailed specifications, rather than high-level generic information models, met higher levels of resistance and poor adoption. As with the early versions of STEP, the mmCif standard developed was extremely powerful, but because it was complicated, it required effort and expertise, generating resistance. ISO 15926 on the other hand, which has been cast at a higher level, specifically targeted ease of use:

“Once groups realize that all you have to do is to produce your own Reference Data Library, they realize, “this is easier than they thought”, and they apply it”.

Several journals mandate submission of data to an identified repository in a format which complies with a specified standard. This is the case for microarray data. Researchers wishing to publish in these journals have to submit MIAME-compliant data (q.v. Tools, section 5, and the provision of tools such as maxload to facilitate compliant submission of microarray data).

Several key informants noted a suspicion amongst some researchers of standards relating to data mark-up. One reported that the term “standards” itself has a negative connotation among researchers. To them the term “standard” means someone is defining a standard by which to evaluate, and criticize the quality of their work and results. They do not understand that metadata standards primarily define metadata used to maintain the data, indicate the applicability of work to the interests and needs of another researcher, and enable experiments to be re-run in silico.

Some data standards are well understood within their domain, but which may be arcane to those outside the domain. The BADC cited some atmospheric data sets as falling into this category; presumably genomic data formats might be just as arcane to atmospheric scientists. Cross-domain data sharing will be hampered by such differences.

Related sections

Metadata and user compliance in supplying metadata
Appendix 5.2 summarizes different categories of
standards (informal, formal, professional, open)

Section 4
Appendix 5.2

7. Planning and management over the data life cycle

- Data plans should be drawn up by applicants at the funding application stage, identifying digital output, data management and mark-up needs and destination (if any) after project end; this should be done in consultation with the funder and informed by funder policy.
- Plans should be reviewed and updated over the course of the project.
- The plan should be drawn up in liaison with the destination repository.
- This predicated the existence of a support framework, such as that provided by NERC or AHDS, which is informed by defined objectives, policy and strategy, through the life cycle of the data.
- Practical support, training and guidance should be provided to researchers, couched in plain English at domain-specific and general level; expertise at data resources such as BADC, EGTC, AHDS, and at bodies such as the DCC and UKOLN should be tapped.
- Awareness campaigns providing examples of scientific benefit through data sharing enhance use of data management support services.
- There should be co-ordination between scientific data management initiatives (such as MIAME, PSI) and data/information management experts (such as BADC, AHDS, DCC, British Library).

This section looks at the life cycle of the data, emphasising the need to plan at the start of the life cycle for quality data and cost-effective quality data sharing in the short and long term. The life cycle of data begins before data creation, and extends through analysis to storage, archiving, preservation and retirement.

7.1 Data planning

To use somebody else's data you need enough information about the data (a) to have the confidence to use the data in the first place and (b) to know how to use it, whether manually or by automated means.

The same applies for the data repository which will hold and manage someone else's data for access by others. Planning and costs are helped considerably if you know what is coming and when.

To avoid difficulty and inefficiencies downstream for all concerned, the information needs to be collected and downstream arrangements for the data need to be identified as early as possible. This planning must begin before data are created. This allows appropriate arrangements to be made for the requisite information to be collected efficiently, in a timely fashion, managed efficiently within the research project and then in later collections.

This activity is eased considerably if standard information requirements have been drawn up, and guidance and tools are provided to support the activity. This has organizational implications (see Section 9).

Planning helps the data producer too. As we have seen above (section 4), quite a lot of information needs to be gathered at data-generation stage. This must come via the producer of the data. Last-minute collection is likely to produce inaccurate or incomplete metadata information. Either these inaccuracies are corrected by curators of a repository or archive to which the data are submitted (but which may not always be possible) or the inaccurate information remains. For those running an archive or data centre, correction of metadata when data are submitted is a very time-consuming but unnecessary exercise which consumes time which could be spent on more value-added work, such as developing tools for users. This point was made by several interviewees.

Planning for the collection of the information supports efficient and cost-effective fulfilment by the data producer and his project team. Tools (such as PEDRo, see PSI case study) can be identified which will support automated collection of the information, the data management and sharing processes.

Planing also alerts researchers to digital archiving issues and should encourage confidence in the use of the digital medium and better data and information management practices at a more general level, such as keeping web links up to date.

7.2 Project resource planning

Data planning at project level also contributes to overall project resource planning. Early identification of likely data management needs fits in with the move to full economic costing in research grant proposals. The same procedures apply equally in the commercial domain:

“Projects where you have not thought about this [data planning] have been “show-stoppers” until this has been addressed ... Where it has not been addressed, the cost to the project escalates, the resource bill goes up...” (key informant from industry)

Preparing this plan allows researchers and project leaders to identify or forecast:

- The digital output of their project
- What work is entailed to fulfil their data submission/data sharing requirements
- What resources (human, software, hardware) are required for this, and when
- Future use and care of the data.

It is very possible that this exercise will allow more accurate identification of data storage needs during the project and identify other needs overlooked by the application’s full economic costing analysis.

Institutions and funders will be in a better position to identify economies of scale at equipment and tool level. Secondly, and more importantly, the digital output of the project which they are funding will be more clearly identified, allowing more systematic asset management, including opportunities for exploitation.

For longer projects, it will be important to review both the plan and its performance.

Recommendation 7.1: Insist on a data management plan as part of the grant application and evaluation process; police the spend; provide incentives to spend the data management allocations on data management.

Recommendation 7.2: Data resources should be established for those communities which do not yet have provision for data retention, in an appropriate organisational form, after an analysis of options and needs.

Good practice

- The AHRC requires funding bids to include a technical appendix **as part of the funding application**, setting out how the bidder proposes to manage the project, including the electronic output, his or her data

development methods, the infrastructure support entailed, and how he or she will address preservation of and access to the output. They suggest where help may be obtained.

The AHRC also specifies that researchers must contact the AHDS within three months of the start of the project “to discuss and agree the form and extent of electronic materials to be deposited with the AHDS”. This means that the AHDS can plan ahead, planning for capacity – people, storage, software resources. This then informs their own understanding, and supports better risk management.

Bidders can and do consult the AHDS team for help in preparing this appendix. AHDS has five specialist sections, so applicants can approach the section relevant to their research, as well as having access to cross-domain expertise. The AHDS report an increasing number of funding applicants approaching them for guidance and has also found that liaison between AHDS and research project continues through the life of the project.

The AHDS guidance and support is easy to find, and contact points are clearly identified. The web pages (www.ahds.ac.uk) are very clearly laid out, uncluttered and easy to read, giving advice on creating resources, submitting applications, and depositing resources. As well as guides to good practice, there are information papers (including an excellent paper on metadata), a number of case studies, and an “acronym buster”.

- ☑ The NERC’s data policy document (sections 5.1. and 5.2, [85]) sets out clearly the issues to be addressed before data collection and the “standards of stewardship” for NERC data. The NERC funded **EGTDC** (Environmental Genomics Thematic Programme Data Centre), as well as providing data management and bioinformatics services to its community, provides cataloguing services for datasets generated within relevant research projects, and may hold the dataset if not held in other public data repositories. The EGTDC provides guidance and recommendations on tools, formats, standards, and can also provide the specialist software.

Under the NERC data policy, datasets “must be fully ‘worked up’ (i.e. calibrated, quality controlled, etc) with sufficient qualifying documentation to be of use to third parties without reference to the original collector”. The EGTDC provides help, guidance and training for this, with information on required standards, provision of software tools for processing the data:

- In the case of microarray data, for example, EGTDC specifies use of the MIAME standard. While the field of metadata is an evolving one, the important point is that metadata to be collected is specified, and the researcher knows what has to be collected, and there are tools to help him, her collect the requisite metadata (see 6 above).

The EGTDC sets out a clear timetable for researchers for the cataloguing and submission of data, starting at project funding stage. This timetable includes review stages, when catalogue entries are reviewed, by both researcher and EGTDC.

The timetable also clearly sets out timings for data release (public access to the data).

- ☑ The NERC data policy stresses the need to address the rights management issues relating to data at the start of projects. (Section 14 discusses rights management issues.) AHDS provides guidance on copyright and licence issues.
- ☑ Both AHDS and EGTDC provide skills-development training. Through its newsletters AHDS boosts awareness of its resources, courses and collections.

7.3 Support framework

Good practice

- ☑ In both of these examples of good practice there is an existing framework to support the custody and stewardship of the data after project end, if the data are retained (see selection section below), and the source of guidance and support is a designated provider within that framework. That framework is informed by the role and objectives of the institution concerned, and by clearly articulated policies and requirements, matched by resources being made available.

Users also benefit from NERC and AHDS expertise and activities in the archiving, preservation and curation (in the wider sense) of data.

7.4 Selection

In some cases research projects will generate data not covered by data submission or publication requirements, or which the PI might not perceive as of possible wider or longer-term utility. For example, in many cases data are generated in experiments or by equipment in a primary state, and then undergoes subsequent processes before traditional statistical analysis or its traditional filing. Other equipment produces thousands of data snapshots every hour.

Decisions will need to be made as to which data to keep. The data plan therefore should be prepared in consultation with the funder or the funder's designated body, and any other stakeholders identified and agreed by the funder, and/or informed by the funder's policy guidelines. Of course, the data will need to be kept somewhere, and will need to be locatable (see NERC EGTDC good practice example in section 7.2 above, and section 15, models). The selection issue is likely to have to be readdressed, however, over the life of the digital object being retained – the reasons or rationale for retention may change or be affected by external factors, such as scientific advance or improved power of instruments (for example, by producing more detailed, better quality data).

The astronomy community is faced with a similar question on selection. One interviewee expressed the wish for multi-tier access to different levels of processing of astronomy image data; stronger and more standardized metadata was needed to bridge the gap between raw and cleaned-up image. Another key informant noted, however, that there was an issue of access by others to the intermediate products created during the process of data analysis, which those generating this data were less likely to wish to share.

At a higher level, the utility and role of some data sets and community resources will remain the same; others, however, may need to be retired as their utility changes with technological and scientific advance. We pick up this point in section 9.

Good practice

- ☑ The Functional Magnetic Resonance Imaging (fMRI) Data Center (fMRIDC) in the USA, managed by computer and brain scientists, “seeks to catalogue and disseminate data from published fMRI studies to the community” [116]. Digital time-course neuro-images of the brain captured by MRI scanners undergo processes in preparation for statistical analysis. The question for the fMRIDC was, which data to keep? Just the final, processed image? It decided to keep data from identified points in the process, to have as complete a set of information as possible for an independent investigator to start with the raw image data and follow the processing steps through to the statistical results [116]. This enormously widens the range of uses of the data – as this referenced article says, it allows a “multiple entry point” approach, offering “the

greatest accessibility across different uses in new science” [116]. Of course, each set of data needs to carry sufficient metadata for the data to be usable.

7.5 Versions and repositories

Many community resources are continuously updated at source and annotation level. Thus research reported in papers is based on data versions which no longer exist, unless the researcher archives a snapshot of the relevant data. How long will the researcher be in a position to keep that data? How long will he need to keep it? Would it be more efficient or cost-effective for a designated repository or the resource itself to keep back versions? A related point is the authoritative nature of data held in designated repositories, community resources (see also data sharing models, Appendix 4).

7.6 Digitisation

In some contexts, such as medical records, planning for digitisation was stated to be important. Digitisation is the conversion of information in printed (or other physical forms); it can simply mean converting to an image (scanning) or extracting information features, such as conversion to text which can be processed (optical character recognition, for example). Issues noted were – what are the costs of this, what are the best technical options, what human resources does it require and given answers to these, how do you decide what to convert into what form.

At the moment there is no single publicly funded, designated place in the UK to get advice on this (there are commercial organisations, such as AIIM⁶). The National Audit Office has drawn attention to this issue, in a wider context than the life sciences research, recommending the British Library take responsibility to take up the question.⁷

Good practice

- The AHDS has an excellent section on digitisation on their web site, giving advice and providing case studies.

7.7 Stakeholders

The increasingly inter-disciplinary nature and horizons of research mean that the potential downstream users and stakeholders in data collections may work in other domains. High-level consultation and co-ordination with regard to scope of objectives surrounding data exploitation are important.

A further point is that some of the data management resources and work to be consumed by the research project may contribute to domains beyond that of the direct funder of the project. This raises funding and organizational issues (q.v. section 9), but it also makes appropriate support and tools to minimize the data management tasks particularly important.

⁶ See: <http://www.aiim.org/>

⁷ See: http://www.nao.org.uk/publications/nao_reports/03-04/0304879.pdf

Recommendation 7.3: Funding bodies should consider cross-domain representation/consultation in data management committees or special interest groups.

7.8 Data archiving and preservation

In several community resources reviewed, long-term responsibility for their holdings (data and tools) was not clear. The resources themselves were set up with other objectives - in other words, archiving is not necessarily their responsibility in the first place - and they lack skills, resources, or assured institutional longevity. Several community resources were aware of the need to tackle the issue, though not all were aware of the technical needs entailed in data retention over time.

Among our case studies we noted that those resources which stored discrete datasets, and for whom continuing high level of **continuous** curation of content was not appropriate, were more conscious of their role as archivists.

Using archived data is a form of asynchronous sharing (as described in Section 3 above). Archiving has additional requirements to those discussed so far. These are:

- Assigning long-term responsibility for data management
- Recording metadata describing long-term preservation needs, the data's lifespan, and information which is otherwise likely to be lost as time passes. The latter includes capturing tacit knowledge which might be generally available now but likely to be unavailable alongside the data (or metadata) in the future
- Demonstrating that authenticity and data integrity have been maintained when disseminated by the archive to future users
- Preserving information, protecting it from becoming inaccessible or unreadable due to media and technological obsolescence.

Models and norms have still to be developed for the long-term management of continuously curated datasets which have no predictable final form or criteria for closure. There is a need to develop appropriate models for these data at organisational, technical and financial levels.

Recommendation 7.4: Resources in some instances (particularly community resources) need to be more aware of their responsibilities with regard to archiving and long-term preservation. We suggest that the sponsors engage bodies such as the DCC in approaching this and take the matter up with the managers of those resources they support.

Setting aside the need for robust models for long-term archiving of continuously curated data, we note a number of specific issues which affect data management and data archiving as a whole and which are still to a large degree unresolved:

Appraisal: the activity of deciding, *a priori* (and at later junctures), what to keep, and for how long. While in a few cases this may not be difficult, in many more cases the ground rules to be applied have not yet been determined. What data do you keep? Who decides? (see section 7.4 on selection above).

Appraisal is closely related to “**disposition**” – archivists’ term for deciding which data to destroy and when. Decisions are likely to be determined differently in separate domains, determining factors being the needs of the science in the domain, perceived continuing scientific, social and economic value in the data, and funding availability. With the increasingly multi-disciplinary drive of research, selection, retention, and the enabling and organisation of discovery and access become questions to be addressed across, rather than exclusively within, domains.

Standards and preservation: Adopting standards which are likely to survive and/or are well described for the future, and whose descriptions remain available makes re-use of data very much easier and substantially decreases the risk of future data loss or heavy expense to resurrect data. Technical questions about how satisfactory preservation of digital objects is actually achieved, and what aspects of data to preserve (content, look and feel, presentation aspects, acceptable information loss on migration for example) remain research issues.

Preserving provenance: Provenance is a component of data trustworthiness to researchers and others. For example, clinicians will be reluctant to use the digital medium unless they can be confident that they will be able to demonstrate their own due diligence, should this be challenged. The issue of preserving provenance in the sense of tracing information back to originals and through workflows is discussed in Appendix 5.1.

- Recommendation 7.5: Funding bodies should liaise more with those researching the question of curating continuously evolving datasets (including the DCC), in particular to determine requirements and objectives.
- Recommendation 7.6: The metadata sets such as MIAME, developed essentially at scientific level, provide a substantial amount of the information needed for long-term preservation of that data. But not all. There should be liaison between bodies responsible for data repositories such as ArrayExpress, Ensembl and experts in long-term data management/archiving.
- Recommendation 7.7: Provenance information is being defined within electronic health record standards. The research community using similar data should seek to adopt the same or compatible standards for this kind of medico-legal and clinical-origin metadata.

Good practice

- The AHDS see themselves as an archive and operate with that as a major objective. Both the AHDS and BADC are adopting the Open Archival Information System (OAIS, ISO 14721, the reference model for digital archives developed for NASA) for their data archives.
- The CDA also noted its archival responsibilities; it outsources its archiving of North Sea oil well exploration data to the British Geological Survey (part of NERC) as part of its regulatory obligations, thereby achieving substantial economies of scale for its members (for instance, storage costs are saved many times over), as well as providing expert care for the data, at scientific and IT level.

- ☑ The Digital Curation Centre (DCC) has been established to provide the research community (and others) with support for digital archiving and curation. It is also undertaking research into these topics. DCC members were involved in development of the OAIS standard.
- ☑ The Royal Statistical Society in collaboration with the UK Data Archive at the University of Essex have produced a good booklet as guidance to researchers on data archiving [100]. An excellent handbook on the preservation of digital materials has also been produced by Maggie Jones and Neil Beagrie, published by the British Library with the support of JISC, the AHDS, and *re:source* (now the MLA). This is currently being revised and updated. [61]
- ☑ In a still uncertain area, one of the key rules of good archiving practice is to keep the original data, as received, as a reference copy. The AHDS do this; it is also the practice the National Archives, the UK Data Archive at Essex University and the British Library, and a few industry archives.
- ☑ Several initiatives recognize that re-running experimental data needs information about the experiment and the conditions under which it was run and which have identified an information set to be gathered. Well-known examples include MIAME and its relative MIAPE. In the case of MIAME, not only is collection of the MIAME data and submission of the data set marked up with the MIAME data a condition of journal publication, but a public repository, ArrayExpress is provided and maintained, in the UK and Europe, by the EBI.

Related sections

Metadata collection	Section 4
Organisational factors	Section 9
Rights on data	Section 11

8. Nature of content

- Data sharing must respect consent and confidentiality of subjects' data
- Failure to do so will result in withdrawal of consent by subjects and refusal to use digital medium by practitioners
- The current administrative complexities of gaining consent for studies involving personal data already inhibit data collection, sharing and re-use.
- To enable further data sharing and re-use, the consent process must take account of potential re-use
- The success of projects such as CLEF, Biobank, VOTES, and the NHS National IT Programme will be critical to the public confidence needed to support sharing of personal data for research

This section looks at issues arising from the content of data rather than its format. These are generally issues of restrictions on use for ethical and security considerations. (Legal and ownership issues are discussed in section 11).

To a large degree these resolve into issues which affect public confidence in the conduct of research. Appropriate management of these issues is a major factor in the quantity, quality, strength (and cost of availability) of data for research and discovery.

8.1 Personal data in research

Of most prominence in this area is the question of the use of personal data in bio-medical and social science research [71]. For data sharing the main issues are of the nature of the consent given by research subjects and maintenance of confidentiality.

Box 5: Consent and confidentiality of personal data

In the context of sharing, the issue of consent is: does the consent given for the original study extend to further use of the information and for how long? If consent is deemed to exist, are there any constraints on the areas of investigation proposed?

Anonymisation or pseudonymisation (where a key is kept separately linking individuals to anonymised data) and depersonalisation (removing identifying information such as names, addresses, etc) may solve, or at least mitigate the problems. (These techniques do not necessarily ensure 100% privacy, as data which remains might, with effort, identify an individual.) The techniques may also actually degrade the data so that it is not suitable for the purposes intended (a simple example – removing address information may make social-class comparisons difficult). Ethical issues remain – suppose someone at risk is identified through the secondary use of pseudonymised data – what action is appropriate?

Obtaining re-consent is another option, but assumes people are still alive and can still be located; it comes with an administrative burden and risks annoying or worrying subjects.

Maintaining confidentiality is a question of good processes and technology, such as having secure systems, good data management practices, and appropriate use of techniques such as security controls, encryption, key management, and as noted pseudonymisation / anonymisation, and depersonalisation.

There is a need for consistent policies regarding roles, access privileges and levels of access to data for researchers.

In the context of data sharing, if personal data are to be re-used in a context other than that for which it was originally collected, this needs to be factored into the form of consent sought. Thus it might be appropriate to gain generic consent at the time consent forms are signed (as in BioBank). An alternative is having to re-seek consent later, or not re-using the data – but this can be costly and as noted in Box 5, subjects may simply not be available to give further consent and the process of getting it may itself be stressful to subjects. Technical solutions are another alternative, such as anonymisation, pseudonymisation or depersonalisation. However these may not be foolproof, and if carried too far may destroy the research value of the data, thus negating the exercise. (see Box 5 above).

For individuals to sign up to broader consent, the success of projects such as Biobank, CLEF, VOTES and also the NHS National Programme is critical.

Good practice

- These issues are critical to the success of Biobank (www.ukbiobank.ac.uk), which will involve the recording of clinical and genetic information from half a million individuals and a follow-through of some 30 years. This promises to provide a role model in the area. The project has set up a governance framework which draws on experts and stakeholders, professionals and members of the public, to ensure consultation is comprehensive. It has drawn up a draft ethics and governance framework to support the project, and has a risk management programme.
- CLEF is a project specifically seeking to establish a framework to support ethical and efficient use of clinical records in research. Building on previous projects, a major effort has been invested into observing strict ethical and legal requirements, working with the Judge Institute of Management in Cambridge. It employs a repository of pseudonymised and depersonalised data.
- The VOTES project seeks to develop a Grid-based framework through which a multitude of clinical trials and epidemiological studies can be supported. Through policy-based security technologies, generic rules, on data access and usage to the various people involved in clinical trials, will be defined and supported to address recruitment, data collection and study management of clinical trials. Primary and secondary care components of clinical trials will both be addressed within VOTES. Other projects are the recently funded Generation Scotland Scottish Family Health Study which intend to link genetics and healthcare.

8.2 Researchers' uncertainty

Data sharing is inhibited where researchers have little knowledge of law or the ethical guidelines, because they tend to be cautious about releasing or re-using data which might lead them into legal or ethical difficulties (particularly if they fear legal sanctions).

The ethical review community may also feel uncertainty. As noted below, the ethical review process is under review in the UK. Ethics committees may have difficulty with the complexity of appraising the risks and benefits of secondary access to personal data, and of defining or judging policies that research groups might adopt. CLEF (as noted below) is not unique in having found difficulty in this area, given the absence of good standards or an accepted code of practice.

Recommendation 8.1: Researchers need better training on both ethical issues and on regulatory requirements and processes. This could be facilitated by provision of clarification and guidance to those funded.

8.3 Administrative burden

The administrative burden on researchers to obtain the necessary regulatory permissions to proceed with collection of personal data is already seen as onerous.

The situation may improve with changes to the Research Ethics Committees' (RECs) reviewing processes and the introduction of Multi-centre Research Ethics committees (MRECs). (Before this change some nation-wide projects had to approach nearly all of the 255 local RECs in the UK individually)⁸. The area is in some flux. Recently the Central Office for Research Ethics Committees (COREC) which oversees the ethical approval process, has moved into the National Patient Safety Agency (NPSA) within the NHS and it is not clear what changes might ensue.

The difficulty may go in two directions; thus the CLEF team noted that though there were problems in their ethical review process, they felt the MREC concerned encountered difficulty because of the complexity of CLEF and the issues it raised, which were more complex than with "routine", local clinical trials.

A multiplicity of regulatory hurdles remains that may need to be navigated, dealing with specific issues. For example, the following approvals may be required: from a REC for general ethical approval, GTAC⁹ for genetic studies, MHRA for drug trials, HFEA for human fertilisation and embryology studies, HTA where tissues are involved and SCAG for the use of national NHS databases. In addition, for certain studies consent may need to be waived, and this is controlled by PIAG.

Given the international nature of much life science research, the problems facing the community in understanding the legal and ethical aspects of data sharing are further exacerbated. Indeed, as outlined in the HealthGrid White paper (at <http://whitepaper.healthgrid.org/>), there is no clearly defined international legislation on data sharing.

An important consideration for medical (and social sciences) research data is that upsetting the subject can cause withdrawal from ongoing or future studies if the subject's rights and privacy are not strictly respected. Thus special care is needed in handling this data, and stricter, but still proportionate controls can help the sharing process overall.

There is a sense that nobody "owns" the problem, though that might be formalised by a grouping of interested research and professional organisations.

Recommendation 8.2: Seek ways in which the complexities of the regulatory processes can be simplified, and the work of those engaged in these processes can be eased, working with other interested parties where appropriate.

Recommendation 8.3: A code of good practice and/or guidelines for good practice should be established for managing and sharing data that is confidential, which recognises both risks and benefits of sharing for all stakeholders.

⁸ As this report was in final draft, a report from the Ad Hoc Advisory Group on NHS Research Ethics Committees ordered by Lord Warner into the operation of the Research Ethics Committees reported (Report of the Ad Hoc Advisory Group on the Operation of NHS Research Ethics Committees. See <http://www.dh.gov.uk/assetRoot/04/11/24/17/04112417.pdf>)

⁹ Acronyms used in the paragraph are defined in Appendix 2.

Good practice

- ☑ The MRC is currently developing practical guidance for researchers to help them understand and navigate through the regulatory process, using on-line techniques; they also produce a set of concise guides to ethical issues and good practice [see also 75].
- ☑ Initiatives, such as those being undertaken by the MRC, Academy of Medical Sciences (AMS) and others are currently addressing the question at both strategic and policy levels. The AMS statement is expected in September 2005.

8.4 Other ethical constraints

We did not encounter any comments related to the use of animals in research and the subsequent use and sharing of data obtained from animal experimentation. However, one informant did confirm that people are reluctant to publish openly the results of *in vivo* experiments on higher organisms for fear of the consequences. The consequence of this is that this information is not generally available and therefore less use is likely to be made of it. Other sensitive areas include stem cell research and genetically modified crops.

Recommendation 8.4: Consider whether guidance for those using data or abstaining from publishing data obtained from animal experimentation needs development.

8.5 Span of confidentiality

Paul Wouters' 2002 report [127] on data sharing policy for the OECD lists the following limitations on data sharing found in his survey of Web documents:

- Need to safeguard the rights of individuals and subjects
- The rights of individuals to determine what information about them is maintained
- Legitimate interest of investigators, for example, materials deemed to be confidential by a researcher until publication in a peer-reviewed journal
- The time needed to check the validity of results
- The integrity of collections
- Data released to the public that could lead to the identification of historically and scientifically valuable archaeological sites [which] could invite looting and destruction
- Data enabling the identification of the location of rare botanical species outside the United States could lead to unwanted bioprospecting and could damage the relationships between researchers and the host community
- Information related to law enforcement investigations
- National security information.

To this list should be added the privileged period of fair use granted to commercial bodies which funded the research to use data following the research project, for a limited time.

The last item in the list has been the subject of heightened discussion in the last few years, with the perception of an enhanced threat from bioterrorism. The need to restrict information, such as the use of biological counter measures, could potentially inhibit sharing. The leader in a recent issue of *Nature*¹⁰ has echoed the need to provide clarity to the community, and to develop clear guidelines.

Good practice

- BADC – in its curation, is aware and takes account of risk and sensitivities regarding rare species and prevention of ecological damage, which can be revealed not just in data but also in metadata accompanying data.
- The MRC issues guidance on the use of animals, including special advice on the use of primates.

Related sections

Intellectual property

Section 11

¹⁰ Vol 435, Issue no 7044, 16 June 2005

9. Organisational and management issues

- The primary focus of community resources should be delivery to users
- Community and data resources should have defined delivery objectives, with matching funding
- Proximity to related expertise brings synergies and cross-fertilization of ideas
- Location of resources within wider organisational framework facilitates efficient management of community resources over their life cycle, from birth to retirement

9.1 Quality of resource – delivery focus

Numerous interviewees reported difficulties simply downloading and integrating data from some community resources, where the cause was poor communication on the part of the resource. Commonly, database schemas were changed without any notification that they had changed, or explanatory documentation was missing. For the user (and supporting computing specialists and help desk teams) these problems resulted in days of unnecessary work, a substantial waste of valuable resources.

While community resources are fundamental parts of life sciences research, by definition one of their primary roles is to support delivery to the user. These resources should have defined objectives relating to the level of service to be provided to users – for instance, defined response times to queries, defined help-desk availability, up-to-date technical guidance. Performance should match objectives. While the service level will be dependent on the funding provided, efficient supporting structures will ease the resource requirement in the short, medium and long term.

Customer-service levels do not have to be gold-plated, but should not be misleading. If a data resource cannot provide reliable “customer support”, then that should be clearly stated by the resource. Interviewees reported that user frustration due to these problems led to decreased use of the resources.

Recommendation 9.1: Resources should have a clearly defined remit and goals; set defined level(s) of service which are achievable within budget constraints; clear policies and procedures; there should be planning (short-, medium- and longer-term).

Recommendation 9.2: Providing clearer separation of research and service roles inside data resources would help foster and maintain more consistent, higher quality delivery of services to consumers, leading in turn to increased level of use.

Good practice

- NASC, BADC, AHDS and CDA all provide efficient and highly responsive service to users, at scientific and data level, with in-house expertise and indirect access channels set up to other expertise, as required.
- NASC, BADC, Ensembl, AHDS and CDA have firm internal management structures and mechanisms, reflected in control over their service provision at any one time and forward planning.

- ☑ NASC (which has a hybrid service role, providing stock grown from seed as well as data), places its priority on service delivery, and thus places a premium on reliability over experimentation, while maintaining development work geared to users and being actively engaged in the research community.
- ☑ Ensembl has a scientific advisory board.

9.2 Research sustains quality

Science and research are, of course, the foundations and *raison d'être* of community resources. Research sustains the relevance of the resource, improves its quality, making the resource more valuable to the science community.

But the research label can be detrimental. Many community resources have funding as research projects. Indeed, a community resource often has to rely on multiple research grants from multiple sources. The negative repercussions are multiple. Firstly, time has to be spent winning research funding and, of course, performing the research. Secondly, it means that the general focus is shifted away from delivering a resource to users, and recognition is meted out not in terms of user benefit but in traditional research measures such as journal citation.

The heterogeneous management and funding context of some community resources can be a diversion. For example, NASC noted that it has to negotiate across three boundaries: commercial, academic and data service, creating extra work. NASC is commercial in the sense of being a supplier of stocks and microarray services, it sits in an academic university department and the university part funds them, but its central remit is to provide data services.

An indirect consequence of the tendency for community resources to carry a research badge, rather than a resource badge, may be that community resources are isolated from contact with knowledge management and information sharing experts, such as librarians, and underlying data management expertise.

Recommendation 9.3: There should be clarity by funders with regard to the objectives of data resources – are they archives as well as active resources, for example?

Recommendation 9.4: Contact and co-ordination should be facilitated between scientists and information/data management experts, such as the library community, digital curation and digital archiving/preservation experts.

Good practice

- ☑ Ensembl, the malaria curation in the Sanger Pathogen Sequencing Centre, AHDS are all examples of resources funded as a community resource, not a research project, and so are substantially less burdened by the administrative requirements of research funding.

Good practice

- ☑ BADC's location on the Rutherford Appleton Laboratory campus allows synergy with CCLRC, and it is involved in many data management projects, and in the DCC.

- ☑ The eBank project, while a pilot, is an example of co-ordination between scientists, information and data scientists, looking at linking primary research data, publications and other information, in both research and education contexts.

Similarly, the AHDS felt that in general there had been little exchange of learning across the arts-science divide. They have many years of experience in data archiving, digital preservation and information management, but they have seen little interest from the science community in what they are doing, except from the MRC and ESRC.

In many cases, we found that there was simply no awareness on the part of researchers and individual community resource managers of issues related to the life cycle of data management (such as digital preservation).

9.3 Location

Location – in physical and organisational terms – can enrich data resources and increase organisational flexibility.

Resources located in proximity to related institutions benefit from synergies - availability of relevant expertise, economies of scale, and from cross-fertilization of ideas. Good examples of this happening are the EBI/Sanger at Hinxton; Manchester, BADC at RAL, and the location of the AHDS subject centres in institutions with relevant expertise. The synergies are apparent at scientific and at an operational level.

Good practice

- ☑ Ensembl is a joint Sanger Institute/EBI project based on the Genome Campus, alongside the Sanger and EBI projects. Not only does it enjoy daily contact with other projects, it can also take part in the EBI's industry fora, and enjoy the specialist infrastructure provided on the campus. For example, Ensembl stresses the key contribution made by the IT systems team at Hinxton in enabling the Ensembl resource to cope with the massive performance challenge the Ensembl resource represents
- ☑ AHDS locates its subject centres in centres with relevant expertise, and maintains a central unit housing top-level and general areas.
- ☑ The Wellcome Trust's malaria *Plasmodium falciparum* curation is done within its pathogen sequencing section, with pathogen curation expertise; this curation work then updated the international Malaria database run from the University of Pennsylvania.

9.4 Internal management

The five case studies running larger-scale community resources – AHDS, BADC, CDA, Ensembl, NASC – all provided outstanding service, based on excellent internal management and use of tight resources. They maintain their resource in the context of very fast-moving scientific advance, which they themselves are also helping to drive. The IT context too is constantly changing. Referring to a resource used by hundreds of thousands of people a year, one key informant compared trying to deal with technical problems and change in management with trying to change a wheel when driving along at 70 miles an hour because you can't stop.

Repository capacity planning can be problematic. It is not helped by information not reaching the repository in a timely fashion. The data planning measures set out in the section 7 should overcome this.

Good practice

- ☑ Excellent internal communication: All five case studies would be able to continue running smoothly in the event of departure of a manager or key individual: at Ensembl, for example, responsibilities are rotated so that staff have experience doing other work; AHDS, though based in several sites, holds regular face-to-face meetings of its managers. A further strength is the strong team spirit fostered at all these resources.
- ☑ All five ADHS data centres are carefully structured into teams, for internal efficiency and improved service to users. The AHDS centralises some of its service provision (such as storage services); ,whilst retaining the separate identities of the five subject data services which are under its wing.

9.5 Organisation of data resources

Our case studies illustrated a number of models for the internal organising repository functions:

Community resources such as Ensembl are of necessity centrally provided, in one or more integrated databases. On the other hand the AHDS and the CDA, which provide a single shop front for downloading datasets, are examples of dispersed provision: these provide a number of subject-specific data stores and services at locations where relevant expertise is available.

For a period, the AHDS had switched structure, but has reverted to a centralised provision for storage and common data management functions, whilst retaining their subject speciality support services for both producers and consumers at dispersed centres within university centres of excellence.

A further model is to decentralise more radically, keeping data for re-use across various institutional and departmental repositories. This can be made to work, but imposes a heavier total management burden and risks loss of control (possibly leading to loss of information). Unless there is a “virtual gateway” of some kind, such as a web portal, it risks confusing potential users and complicates their discovery work. Further, some support functions, such as those dealing with IT questions and preservation, are probably best done centrally whilst those which are domain specific are best done close to the domain community.

Another organisational model to mention is *openEHR*, an international, not-for-profit organisation, still young, one of whose principal aims is to help achieve inter-operability in electronic health records. It does this, *inter alia*, by participating in standards development and by developing open-source specifications, software and knowledge management resources for electronic records. These resources (not data) are prepared by leading experts with many years of experience in the field, and are freely available to all. *openEHR* has been set up as a not-for-profit foundation. Awareness of the resource internationally ((thus availability of some monies for outreach) will be critical to its success.

Whatever model is adopted (and no one over another is recommended here), it must supply basic requirements through the data flow from producer to consumer. These are summarised in section 15.

Good practice

- ☑ The NERC DataGrid work is addressing the potential hindrance to cross-domain work of a distributed data store serving many (sub)disciplines. So too is work on the semantic web (see also CLEF case study, Appendix 3.4).

9.6 Co-ordination across subject domains

There is a need for a communications vehicle/channel/pointer to resources, for the different practitioners and players involved in data planning/curation/management. Example: metadata fields (MIAME, NERC DataGrid). This will help avoid re-engineering of metadata structures.

Related sections

The desirability of early data planning and adherence to the plan.	Section 7
Costs that should be assigned to data management	Section 15
Need for data resource career structures	Section 10

10. Culture and careers

- Both funders and individuals are key to a sustained, open culture of sharing
- Hostility in the community will be avoided by sensitivity to individual circumstances when setting periods of privileged access
- Recognition of (and mechanisms for) data citation will promote data sharing
- There should be recognition of the role and skills set of data resource managers; promotion boards need to be briefed on the importance of the particular needs for data curation/management/archiving skills, and which do not fit established career categories.
- Data management and curation accreditation schemes need to be developed
- Data citation and contribution to repositories need to be incorporated into criteria by which researchers are judged. This in turn will provide incentives to users to participate in data sharing actions (such as metadata provision), and foster awareness and skills.

10.1 Community factors

Many of the resources reviewed during preparation of this report exemplify an open culture of sharing and access. Key factors for an open culture are the role of funders and of key individuals, prevailing in particular over perceptions of potential professional or financial advantage arising from a period of privileged and therefore closed or restricted access to resources.

The *Arabidopsis* community is one such example. The plant itself has no crop value, which was an important factor in its adoption as a model organism in the early 1980s. The absence of such a resource had been an obstacle to research.

The human and malaria genome sequencing projects are model illustrations of open access to resources. Because of the firm stance and persistence of funders such as the Wellcome Trust, and key individuals, these resources are now available to all, free of charge. For both, the journey to achieve this was one of considerable work, persistence and some tension, asserting the continued open status of the resources.

The malaria *Plasmodium falciparum* sequencing (see case study) established a formal structure for the project, with clear communications channels, ensuring community consultation and participation at all stages of the project, and going to substantial effort to ensure that researchers in emerging countries (where, of course, malaria is endemic) could participate. In the early days of the internet, this was not easy.

10.2 Sensitivity to circumstances

Key informants also pointed to the need for sensitivity to individual factors and circumstances. For instance, the size of a laboratory or unit should be taken into account in the assessment of privileged period of access. Simply divulging a data request can provide a pointer to other researchers in the same area of investigation. While this can be insignificant for a well-endowed laboratory with a large staff and lots of equipment, able to start at will and concentrate resources, that is not the case for the small unit already doing the research but with thin resources: it may find the large laboratory harvests the rewards of recognition and publication first. Any criteria for data release dates should take

account of individual circumstances; interviewees pointed out that lack of sensitivity would generate hostility and resistance to related data sharing initiatives.

Another key informant pointed to individual confidence and seniority as factors affecting propensity to share: a young post-doctorate researcher, with his or her career yet to build, is much less likely to have the confidence to release his research data than a senior researcher with many publications under his belt. Nor is seniority the only factor: he noted some people are innately more confident than others, and correspondingly more likely to release their data earlier. This would also seem to be borne out by a survey conducted in 2000 [34] by health-policy analysts at Harvard Medical School, which found that

“young researchers, those researchers who publish a lot, and investigators seeking patents are most likely to be denied access to biomedical data and reagents.”

The article continued:

“[the study] also found that researchers who withhold data gain a reputation for this and have more difficulty in obtaining data from others”.

Reluctance to share is particularly understandable where the data are one of the researcher’s primary outputs, and the result of a large amount of dedicated work. A prime example is that of population datasets, which sometimes represent a substantial part of a researcher’s life work. Critical to such researchers’ support for data sharing will be (a) confidence in the continued integrity of their data set, (b) confidence that their authorship of the data set will be recognized in all use of the data set, (c) an organisational framework, mechanisms and tools which support (a) and (b).

The problem, whatever its facets, contains an element of game theory: an environment of reciprocal sharing means everyone has access to the benefits; however a defector can upset this balance for personal advantage, at least in the short term. Several key informants made similar points: one feared that sharing might be (or perceived to be) a one-way street – the effort will not be reciprocated. Two others pointed to the delicate balance which open data sharing represents: one selfish individual (whether data producer, user or resource itself) could easily compromise the open culture.

10.3 Fear

Fear was frequently cited as an obstacle to data sharing. Foremost, perhaps, was the fear that release of data will enable somebody else to “steal” your ideas or beat you to publication.

Another fear is lack of confidence in your scientific work – that somebody else will find an error, disprove my thesis. Of course, this is one of the reasons why data should be shared, though correction should be handled sensitively.

Another major fear is misrepresentation of data or work, or distortion of data and/or findings. This is particularly true for those who have invested a particularly large amount of time and effort in collection of the data.

Fear of loss of control over data is another major disincentive; this came in several guises, another being worries about corruption of data by hackers or inexperienced users.

Good practice

- NASC allow a period of three or six months' exclusivity on data provided by those donating stocks, but link this to the priority with which they will be serviced during that time – the shorter the period of exclusivity requested, the higher the position in the queue for services.

10.4 Recognition - producers

For researchers who generate data within their research project(s), their primary objective traditionally has been the publication of a paper in a journal, in as high a quality journal as possible, increasing their citation count. The same applies to the funder of the research. Until recently, deposit of data barely figured in this career paradigm.

Today, an increasing number of journals require data deposit as a condition of publication, and this may contribute to an adjustment in the career paradigm.

Preparing data for sharing, in particular supplying metadata, is not a rewarding task for researchers, since it is not work for which they receive any immediate benefit, and requires time which could be spent on tasks seen as more relevant to careers.

The net result of this is that data sharing itself is compromised:

- The discoverability and discovery of data is impaired, possibly severely, both in terms of quantity and quality
- Even if found, the data may be much more difficult, if not impossible to use because of lack of information as to how to use the data
- The information provided may be inaccurate, leading to misinterpretation or misapplication, impairing the quality of work
- Insufficient and/or inaccurate metadata about data increase the cost and risk factors entailed in preserving the data over time.

All these risk discouraging people from making secondary use of data.

A number of informants made suggestions for measures to encourage researchers to provide better metadata, listed below:

- Providing a period of exclusive access to data for the producer once deposited, that they can more fully exploit it during that time without fear of competition.
- Include data citations in research publications and encouraging project review committees to give research credits for data citation.
- An argument used by the AHDS to data depositors is that they themselves might want to access the data in the future and will need this metadata

Good practice

- Germany's National Library of Science and Technology now acts as an agency registering Digital Object Identifiers (DOIs), in cooperation with the World Data Climate Centre at the Max Planck Institute for

Meteorology and others. The pilot project has registered over 175,000 data sets (March 2005), and expects to register one and a half million or so by the end of 2005. These DOIs facilitate data citation.

- Many journals now provide references to where data or other supporting information can be found (e.g. Nature). There is increasing use of using persistent digital identifiers such as DOIs for their articles.

10.5 Career paths - service providers:

The activity of providing a life-sciences data resource does not correspond to any existing profession, so terms are thin on the ground and unstable, and there are no career management structures.

What incentive is there for an individual to enter or stay in a job which offers a poorly defined, low-profile career path? Promotion committees have little knowledge or understanding of the function, the difficulties of the job and the expertise involved. The activity falls between several stools – IT, domain expertise, service management - so it is difficult for traditional personnel departments or promotion committees to see where this function fits, and identify appropriate assessment criteria accordingly.

Further disincentives are less tangible: Service provision in any area tends to be taken for granted when working, but it is criticized heavily when there are problems (possibly exacerbated when the service is provided free). Lack of respect is another discouragement: some fields of data service provision require knowledge of a wide range of topics and data sets – “using a river analogy, you need the broad and shallow rather than narrow and deep” – so a service provider may work in an institution or a community where depth of knowledge is one of the main factors for respect and advancement.

On the other hand, relative to a short-term research post, a full-time service provider post offers substantial attractions in employment stability terms.

Recognition of the role also results in tension between the academic (university) pressure to do research and publish, and the desire to maintain an open, service orientated organisation..

A number of groups interviewed noted the lack of professional career structures for data managers, and curators. Nor were there recognised professional qualifications.

BADC noted that promotion committees have a poor understanding of the data management function – it is seen as neither a scientific nor a IT function. This, allied to the lack of career structure, leads to staff recruitment and retention problems.

Recommendation 10.1: There should be mechanisms (tools) to support due recognition of the work/contribution of individual(s), and mechanisms to support rights management relating to data, where appropriate, over the life of the data. This will encourage data generators to participate in data sharing.

- NASC makes good use of the pool of highly trained and qualified scientists who are mothers who want stable, part-time employment which makes use of their skills.

Related sections

Training and awareness for researchers

Section 11

Large-scale data sharing in the life sciences

11. Support, training and awareness

- Provision of good-quality support materials written in accessible language is vital to encourage participation in data sharing tasks by providers and consumers
- There is a role for local, “on-the-job” training by colleagues, as well as support from data resources and external bodies.
- While some bio-scientists are computer literate, some sections of the community may still need further training opportunities, and incentives to make use of them.

Training is essential to quality sharing of data, at all stages and levels. Training is needed to help the data producer collect data and submit it with the appropriate supporting information to destination repository; training is needed to help users use the data and tools, and training extends the skills of those providing and managing the resources.

Training in basic and higher computer skills will enable scientists to make fuller use of data and tools. As one key informant remarked, many researchers are far from making full use of the power of computer-based statistical or visual analysis applications, which would be remedied by training. Basic computer skills would provide scientists with access to powerful tools which have not been engineered to work with a web interface. Another key informant, stressed also the value of learning by hands-on use and turning to colleagues for help.

We saw many examples of good practice of giving support the user community, whether producers or consumers. These included:

Good practice

- ☑ Provision of a help desk: At BADC this is a central function, as essentially it deals with one discipline; within the AHDS help is provided through the five subject–specialist areas (with central support for generic matters). NASC has a full-time member staff designated as a Customer Service Manager. See <http://ahds.ac.uk/> for the AHDS and <http://badc.nerc.ac.uk/home/> for the BADC.
- ☑ NERC’s Environmental Genomics Thematic Programme is also exemplary, and a model of forward-thinking provision for users, providing training on key tools and technologies as well as access to these and development of generic tools tailored to the field of environmental genomics.
- ☑ Imaginative use of web-based support resources for scientists is made by the NCRI in its National Cancer Informatics Initiative (NCII) in showing what is available and the state of play over different specialities (DNA to epidemiology) related to different technologies and infrastructure elements. The NCII’s vision includes promoting data sharing and standardisation in informatics. See <http://www.cancerinformatics.org.uk/>.
- ☑ The use of a web site to provide assistance. In some cases this is extensive. AHDS, and BADC and its sister data centre the BODC are excellent examples. Notable is the AHDS, providing extensive guidance on metadata, exemplars and case studies, reference to useful articles, areas for those specifically for those creating resources, those depositing resources, and those searching their collections. These are written in clear, plain English, and contact addresses and telephone numbers are clearly provided. The BADC in their help section provide extensive information about individual data resources, and ease of access to download data. In a commercial context, the CDA provided similar support.

The AHDS remarked that they now see people approaching them when these researchers are starting up a project, providing an encouraging measure of success.

- ☑ We noted that the BADC has a group which helps projects draw up data management specifications.
- ☑ The AHDS hold regular workshops on specific topics, such as digitisation.
- ☑ Several case studies and related units provide newsletters and run mailing lists, all providing useful pointers to new resources (examples: AHDS newsletter, the GARNISH newsletter in the plant community).
- ☑ Provision of handbooks and manuals. We noted the CDA's manuals, the Royal Statistical Society's archiving handbook, The British Geographical Survey's handbook on the management of metadata, and the MRC's handbooks on research practice and ethics.
- ☑ Community resources with a large, world-wide community of users face demands of a different order of magnitude. Ensembl gets tens of thousands hits per day; as other EBI resources, it provides online and down-loadable tutorials and walk-through examples of using the site.
- ☑ Ensembl and the CDA's web design was informed by feedback from users and marketing analysis.
- ☑ Professional societies play an important role in informing the broader community of resources; in several cases, initiatives for resources were taken forward through the agency of the professional society – for instance, the Society for Neuroscience, behind the Neuroscience Database Gateway initiative.
- ☑ IT training on a commercial basis can be very expensive; the BADC has collaborated with other groups in the RAL where they are sited to run training more cost-effectively.

Related sections

Planning for use of data

Section 7

Career issues

Section 10

12. Legal, regulatory and rights management issues

- Legislation can act both positively for, and negatively against, data sharing. More care therefore should be taken in framing and informing legislation before its entry into law.
- Use of “non-viral” open source licensing is to be preferred over the GNU style of licence

Legislation and data sharing

As we noted in chapter 8, data sharing in medical research in particular is hampered by uncertainties about the scope and applicability of legal and regulatory framework around obtaining consent for data (re)use and maintaining appropriate levels of confidentiality.

Legislation can promote the availability of data – in intent, the recent introduction of the Freedom of Information law is a case in point. It can also inhibit the sharing of data, by restricting access to materials: for example the European Database Directive enabled copyright to be filed on compilations of facts, even on unoriginal compilations. This raised the risk that a company might copyright data needed by the research community, condemning the community to cost of access. Professor James Boyle of the Duke Law School [Financial Times, 22 November 2004] argues strongly that the database directive is “drawn too broadly and triggered too easily” and points out that “the USA added many more databases without a database right than the EU did with one” (meaning that the USA created more databases for third-party use than the EU).

Hasty legislation to promote economic growth through corporate innovation by providing protection to intellectual property can have precisely the opposite effect, stifling access to core research materials. Legislators should take into account the key importance of what is now often called the “creative commons” and, as James Boyle advises, conduct thorough research and analysis before introducing laws such as the database directive.

On the other hand, it was this very fear, of patents restricting access to resources such as the human genome sequence, which spurred the creation of community resources which hold data which is a public good.

As one key informant pointed out, data sharing raises the issue of protection of intellectual property more sharply than was faced with publication of papers in journals. This is because intellectual property tends to reside visibly in the content and nature of the data itself, whereas the publication can say what it wants to say. A disincentive to sharing relevant data for anybody wishing to exploit the intellectual property, this can be addressed by the time granted for privileged access to data.

Good practice

- The human and malaria genome projects are examples of resources which are publicly available thanks to funders and the efforts of individuals, acting selflessly for the good of research. NASC pointed out that the use of Arabidopsis as a model organism arose in part because of this issue, as Arabidopsis has no commercial crop value.

Rights management

In many cases researchers have invested a great deal of effort, care and deep commitment in data sets they produce. This is particularly common with epidemiological and field data: the data sets themselves are the primary output of the research, alongside publication, rather than a by-product. However, as the OECD Working Group on Issues of Access to Publicly Funded Research Data stresses as the core principle, “publicly funded research data should remain publicly available, subject only to compelling superseding considerations and policies” (Peter Arzberger, presentation at the 18th CODATA Conference, 30th September 2002). It is therefore critically important that (a) all due acknowledgements are made on any and all re-use of these data sets, (b) the data sets are re-used with respect, and that any changes or adjustments should be shown, and where possible agreed with the producer and owner of the dataset. For this to happen, rights information must accompany the dataset. This implies that as people move on and even institutions move and change (which the rights information must take into account) rights can be respected despite such changes over time.

If data producers and funders can be confident that their work and role is recognized and respected, they will engage in the data sharing process. Data citation mechanisms are only part of the answer. Time-proof rights information, we repeat, must accompany the data.

Furthermore, however, the data sharing framework and mechanisms must be able to confirm the integrity and authenticity of the data set cited. This complex question (related to the issue of provenance, which we present in detail in Appendix 5.1) is being addressed by bodies such as The National Archives, whose responsibility includes the ability to demonstrate the integrity and authenticity of a record in court. As one key informant stressed, provenance, data integrity and authenticity are absolutely critical for clinical practitioners: unless they can demonstrate due diligence in their care using the digital medium, they will not use it. To meet these needs requires documented operating procedures and processes, and data architectures to support the additional information needed (such as workflows, audit trails, and extra security features).

Good practice

- The Open Geospatial Consortium is conducting a digital rights management project, consulting governments, businesses and academia, to inform the development of open standards for geospatial data and services. As its October 2004 call for input stated,

“Digital rights management .. is a technology for describing, identifying, trading, protecting, monitoring and tracking all forms of rights usages, including management of rights-holder relationships. ... The current inability to control the flow of such information activities has been a barrier to broader adoption of web-based geospatial technologies. Some government data has been withheld from public release over the web because of an inability to manage owner rights after the data are released.”

- Data is more likely to retain its integrity in a designated repository or archive than when held by individuals or units without awareness of and/or means to maintain data integrity; conversely, retention in a repository or archive (such as the BADC, AHDS, which are implementing OAIS, ISO 14721) which demonstrates due care of its holdings, day in, day out, will itself confer a degree of authority on data held.
- The Digital Curation Centre has an important research programme looking at issues and solutions to demonstrate provenance in data sharing.
- The Lambert Review of Business-University Collaboration highlighted a lack of clarity over IP in research collaborations as a major barrier to business/university collaboration. Following the review a very clear walk-

through site and toolkit has been set up on the DTI web site, guiding universities and publicly funded institutions on how to approach and assess IP aspects of possible technology transfer projects: Toolkit url: <http://www.innovation.gov.uk/lambertagreements/> and see also: <http://www.patent.gov.uk/about/ippd/knowledge/lambert.htm> .

The importance of open-source licences

Many of our informants identified the critical importance of the use of open-source licences in developing tools for scientific data sharing. Open-source licensing means that:

- projects can rely on the software continuing to be available and in worst case could provide their own support for it.
- communities of users can share the support and development costs
- specific requirements/modifications can be made to the software on the timescale of the project (at the project's expense of course).

It was also pointed out that GNU-style licensing, where developments in the computer code must also be made open source, was often inappropriate because it hampered industrial collaboration, partly because of the difficulty of being able to exploit the end result commercially, but also because of the legal complexity associated with such “viral” licences. Companies are concerned about the danger of getting into situations where they might lose control of their existing IPR.

Related sections

Tools development	Section 5
Archiving and preservation	Section 7
Nature of content and questions of consent	Section 8

13. Corporate sector

- Many of the issues faced by the commercial sector are the same as for publicly funded sector, but they do have the discipline of the market to focus their efforts and to collaborate where there is little competitive loss.

Data sharing in the corporate sector

Businesses face similar data sharing issues at technical and cultural level as publicly funded research, but they can address many of these within less diffuse boundaries and with the financial incentive of higher productivity (in terms of quality as well as quantity) which results from interoperability of data and computer systems. We take for granted that competing banks share data with each other – it is a *sine qua non* of financial transactions.

Less high profile, but more challenging in technical terms, are the data sharing activities conducted between companies operating in technical areas such as engineering and pharmaceuticals. Both are vertically integrated industries, suggesting not only that the consecutive data sharing defined in section 3 above is common, but also that data traverses multiple transition points (so difficulties in exchange and working with data are repeated at each stage). A prime illustration is given in the CDISC example at the opening of section 6 above. Key informants from the corporate sector noted that, in the broad petroleum industry, between 0.5% and 3.0% of total capital cost of a project (often costing many billions of pounds) “goes on getting information into operations” – that is just the engineering information. They referred to the large number of studies into the amount of time spent by an engineer spends looking for information and data, some of these estimating that this can take up to 60% of an engineer’s time, leaving 40% for productive work. The incentives to achieve ready exchange and sharing of data were and are great, and led to initiatives such as STEP and CDISC.

The Common Data Access, ‘CDA’, case study is an example of competing companies coming together to set up a joint data sharing facility, for economies of scale and improvements in quality and efficiency. In CDA’s case, the data are oil well data relating to the off-shore continental shelf in the North Sea. It has some 25 fee-paying members, who have an obligation to keep certain data in perpetuity. CDA also runs a data registry open to the public, on its web site.

CDA users estimated that common storage provides a five-fold of savings in storage and back-up costs – instead of holding a proliferating number of copies of data, each slightly different, in different places in any one company, each having to be backed up, one copy of the data is held in the repository, which itself is backed up by a specialist data storage and repository company. They did note, however, that storage costs are becoming a conspicuously heavier cost item.

The CDA, like the AHDS, has a layered structure. Actual storage and back-up of the data itself is managed by a third-party company, at lower cost – in the same way that the CCLRC’s Atlas DataStore can purchase equipment in bulk at much more favourable rates than a small unit. The British Geological Survey, one of NERC’s data centres, provides data and data management, while CDA, a not-for-profit company with DTI funding, provides overall management.

As with data sharing in publicly funded research, consultation and co-ordination between companies is essential. For example, CDA works very closely with its Datastore user group. User consultation

reveals that users access data in surprising ways, or use data in unexpected ways, and means that CDA can then adjust its own provision for this.

Good practice

- DTI funding of standards work such as ISO 10303 (which accounts for some two thirds of ISO activity).

Data sharing sustains mature business areas

CDA management pointed to indirect benefits of data sharing, beyond cost savings. The North Sea is a mature oil and gas field, where the major oil wells have already been pumped. By providing access to oil well data, operators can identify commercial opportunities which collaboration make viable. By comparison, Kazakhstan is a much younger oil area, and there is very much less sharing of data.

Between companies, commercial pressures apply – supplier companies having to submit data in accordance with the customer’s specifications. With pressure from customers and the increasing advent of open standards for data formats and operating systems, proprietary lock-in is becoming less of a barrier to data sharing.

Data quality

The CDA pointed to data quality as a major issue for them. The data they received into their repository does not always arrive in optimum condition. Project managers had no incentive to provide marked-up data, even if they had time to do so. Sometimes elements are missing. All these points echo findings reported by BADC and AHDS. For the CDA and BGS, this means spending time identifying and remedying the gaps in data and metadata.

Again, other similar problems arise: they have had to work with many different types of data, which uses different nomenclatures. There were differences at very basic levels, such as the scope of definition of an oil field. They found they needed to identify data models and establish a formalized data schema process.

Two of the large management consultancies, PricewaterhouseCoopers (‘PwC’) in 2004 and Deloitte Touche in spring 2005, have each conducted surveys into data quality. The PwC survey reports that over half of survey respondents said that sharing of data to third parties is likely to increase over the next three years. At the same time, however, while respondents are at least “somewhat” dependent on third-party data, only 18% of respondents whose respondents do share data with others are very confident in the quality of that data; 24% report little or no confidence in third-party data.

The primary barriers identified by those not confident they would have a formalized strategy for data quality included:

- Budget
- Senior executive support and appreciation of the benefits
- Knowledge about how best to manage data
- Belief that there is a problem.

Data quality – value and cost

Respondents to the PricewaterhouseCoopers survey were also asked to define the value of data. On average, respondents estimated that the value of their data at some 37% of the worth of their organization. Only 15% of these organizations actually calculated the financial value of their data, and, among these, no single method predominated.

We believe this figure of 37% presents an interesting correlation with the estimates provided by key informants working with publicly-funded research data as to the percentage cost of the management of that data over the life of the data. These ranged between 10% and 30%. An article in *Nature Bioinformatics* in September 2004 (“Funding high-throughput data sharing”, Vol. 22, pp. 1179-1182, by Catherine A. Ball, Gavin Sherlock, Alvis Brazma) estimates the cost of management of microarray data, from collection through repository life, at 25%. While dealing with data generated by complex laboratory processes, his estimate does not take into account costs for archiving/preservation of the data. This figure underlines the importance of tools and standards to minimize unnecessary effort and loss of productivity, and their corresponding benefits of higher quantity of data and higher quality.

Corporate use of community resources

Business use of environmental data held by NERC data centres such as the British Geological Survey and the BADC is substantial. An analysis of net costs showed that charging for the marginal costs access would have exceeded the revenue generated. This point and the economic value of availability of facts and ideas - “the public roads of our information landscape” - is developed by Professor Boyle (*Financial Times*, FT.com site, 22 November 2004).

Equally substantial is the life sciences industry’s use of data and services provided by the European Bioinformatics Institute. These are provided free of charge. However, 18 large life sciences companies located in the UK are fee-paying members of the EBI Industry Forum: their fees cover events and training, but also provides an opportunity for members and EBI to discuss issues, themes and needs. The EBI also runs a similar forum for small and medium-sized businesses, providing interaction between businesses and with the EBI, and an excellent communications channel of business needs to the EBI.

Life sciences companies contribute funding and resources and participate actively in other community-resource initiatives such as the SNP Consortium and now the HapMap project, both also funded by the Wellcome Trust.

At the end of last year pressure on pharmaceuticals companies following several cases of “adverse events” (damaging side effects caused by drugs either commercially available or in clinical trial) led several to set up web-accessible clinical trial data registries, publicly available.

Data sharing between business and publicly funded research

Several key informants stated that there was a lack of confidence on the part of companies in the security of communications and computer networks in publicly funded sectors. They confirmed that this was a frequent cause of companies deciding not to enter collaborative research projects where highly confidential data and proprietary applications would need to cross these publicly funded networks.

Related sections

Data planning

Section 7

14. International issues

- In the future coordination on an international scale will become more pressing.
- Policy statements at national level should be translated into policy in practice.

14.1 Internationalisation of science

It has been noted above that contemporary technologies break down international as well as subject boundaries, and this is particularly true of data sharing, where scientists on a global basis deposit and retrieve data irrespective of boundaries, and irrespective of who funded the production of the information. This development is likely only to increase, with emerging tools for providing for establishing “virtual organisations” to undertake specific projects, enhanced collaboration, and access to remote instrumentation and sensors. Grid and similar technologies underlie many of these developments. In some instances the cost of the equipment and computing resources needed (such as high performance computing) to solve the most complex problems is beginning to outstrip the capability of all but the largest national economies, driving forwards further collaboration. The organisational and other implications of this are only just beginning to be examined [4].

However, while at an operational level this is happening within the scientific community, at the moment there is little in the way of a consistent approach to many of the issues at a higher, policy level. Such areas relate to differences such as intellectual property rights, policies with regard to openness and data deposit, privacy and funding.

In January 2004 the OECD Committee for Scientific and Technical Policies issued a statement at ministerial level of the principle that publicly funded research data should be openly available¹¹. From such a declaration the coordination needed may (slowly?) emerge.

While many practicing scientists will be unaware or unconcerned with the policy implications of these developments, working over an international scale is becoming part of the set of assumptions that they bring to their every-day work lives. The NASC case study demonstrates the complexity of the international connections within which plant scientists are working. The countries represented by the initiatives and organisations shown on the diagram in the case study report include the UK, USA, Germany, the Netherlands, Belgium, Spain and France. Funding will come from all these countries, as well as through international bodies such as the EU.

Because research is an activity which takes place on an international canvas, whether in the sciences, social sciences or humanities, the infrastructure for research is built on components which are international in context. This is no more so than in the standardisation and ICT area where generally developments can only be maintained and asserted in an international context. This is as much true for semantic standards (field labelling, vocabularies and ontologies, for instance), as for computing interoperability and metadata.

Relevant international groups which should be mentioned in this context are: IFLA, the International Federation of Library Associations, is working on metadata; W3C is working on web services which

¹¹ See www.oecd.org.

are now being taken up by the GRIDs community. The Global Grid Forum is working on GRID standards in middleware and for various kinds of information access / exchange.

14.2 Initiatives

In the USA the National Science Foundation (NSF) is heavily funding the Shared Cyberinfrastructure (SCI) initiative¹². This initiative supports, among other things, high-capacity mass-storage systems, system software suites and programming environments, productivity software libraries and tools, and large-scale data repositories.

The Australian government has recognised the need for establishing a strong e-Research base, and through its “Backing Australia Ability” strategy of 2001 and 2004 has committed Au\$8.3 billion to develop the country’s innovation base over a ten-year period. As part of that, Au\$1.0 billion has been committed to a modern e-research infrastructure.

A European Research Council (ERC) is in formation. Many UK groups working in the area receive substantial funding from EU sources

Though these examples such as these are not wholly related to data sharing these initiatives will have an impact through developments in infrastructure, in the wide sense discussed in section 4.4.

14.3 International data sharing

There are examples of good data sharing on an international scale over and above those described elsewhere in this report, such as NASC, BRIDGES and Ensembl. The Global Biodiversity Information Facility (GBIF) is another good example of sharing information internationally. The project was established in 2001 to make it possible for policy and decision-makers, research scientists and the general public all around the world to access electronically the world's supply of primary scientific data on biodiversity¹³. GBIF is coordinated by a small secretariat located in Copenhagen, and brings together diverse resources from 132 repositories from many countries.

Outside the life sciences area we draw attention to just one example, the International Dunhuang Project, which has made images of ancient documents and supporting materials found along the Silk Road available for scholars and the public. It is backed up by an excellent on-line resource in both English and Chinese¹⁴.

Good practice

- GBIF – knitting many resources together over a wide range of countries and data types
- The International Dunhuang Project demonstrates a scholarly collaboration to make information available across different languages and cultures.

¹² See <http://www.nsf.gov/div/index.jsp?org=SCI>

¹³ See <http://www.gbif.org/>

¹⁴ See <http://idp.bl.uk/> This is made available in English through the British Library.

Related sections

Infrastructure

Section 4

Tools development

Section 5

Standards setting

Section 6

PART 3: CONCLUDING REMARKS

15. CORE MODELS FOR DATA SHARING

15.1 Funding

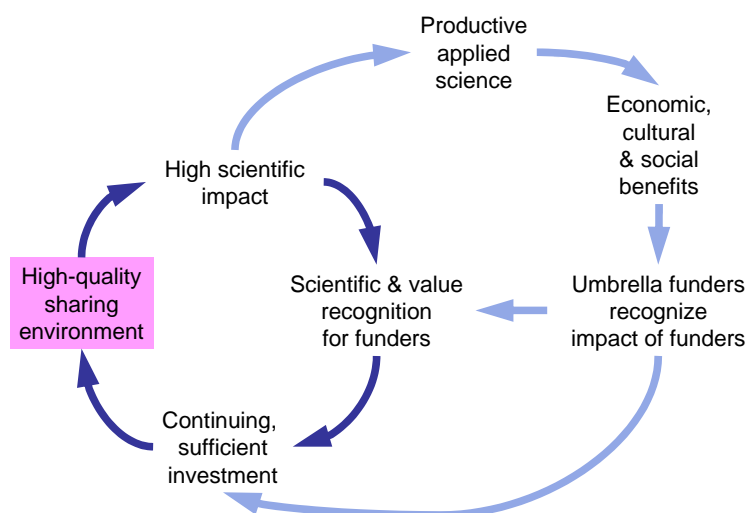
Models of effective data sharing must address and underpin several perspectives:

- The objectives, needs and responsibilities of the various players in the data sharing cycle
- The desirable characteristics of sharing processes
- Organisational structures
- Funding
- The wider contexts in which sharing takes place of scientific communities, international aspects, the various IT, professional, standardisation infrastructures which support sharing.
- Professional reward structures

From the study's findings, we list the following features of a core funding model for quality data sharing, covering data held in both centralised and distributed structures. We believe these are investments in quality and routes to cost savings, not just in data sharing, but in overall IT equipment and management productivity. Other benefits key informants highlight are increased level of technology transfer, both at science and computer software level. For several items listed, initiatives are in progress.

1. Stable public funding for major resources covering housing of the resource, research and development for sustainability, maintenance and integration with new resources, and above all, sufficient funding with appropriate staffing for the resource to meet clearly specified delivery objectives.
2. Allowance within individual research projects for data management ahead of possible downstream sharing of digital output (data, metadata, workflow, processes, tools), beginning with planning at project application.
3. Funding for training of researchers for this activity.
4. Funding of domain-specific support for this data's management, from planning through to submission and ingest into repository or resource; this domain-specific support should share expertise, tools and possibly also infrastructure where appropriate. Though still young, excellent examples of this are the NERC EGTDC and AHDS.
5. Funding of tools to support the data capture, collection and downstream management process, including repository tools, digital rights management, selection, provenance.

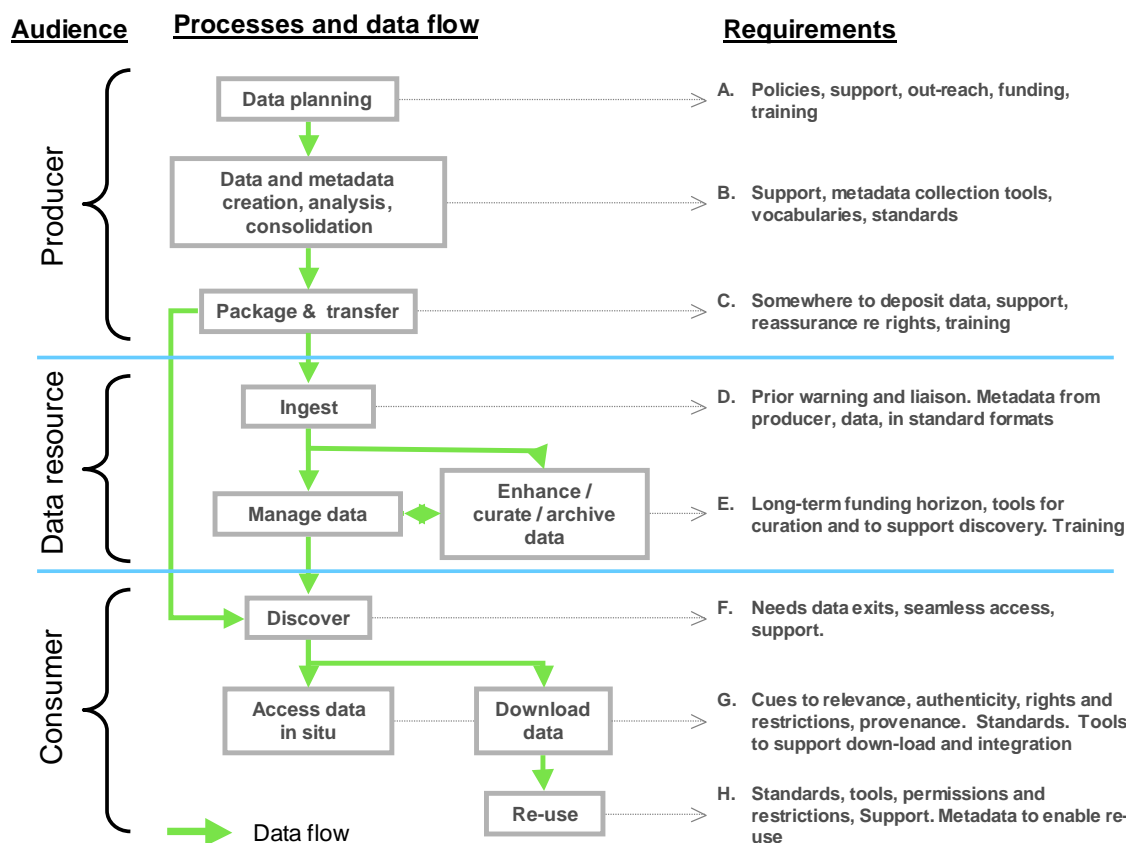
6. Funding for repositories for data which should be retained but not destined for a major community resource, and cataloguing mechanisms to ensure broad discoverability.
7. Funding for work on development of standards, and their implementation and encouragement of adoption.
8. Funding of tools for more powerful data acquisition by users – searching across multiple data types held in multiple locations, and mining complex data; funding for tools to leverage data.
9. Funding for a range of computer training, from elementary to more advanced, to enable scientists to derive more value from data and data-analysis tools.
10. Funding for an infrastructure supporting federated access and federation of data.
11. Funding for resources to support quality software development (for example, a single code repository or integrated code repositories per Research Council) and training for informaticians and computer scientists. The OMII institute also supports this need.
12. Funding and support of active interaction and collaboration between bench scientists, informaticians, computer scientists and knowledge and information experts. Contact between those with digital archive and preservation expertise and scientific data management experts is crucial for the sustainability of the data sharing base, and avoidance of duplication or wasted work, in particular at metadata level. Projects such as e-Bank, linking data with publications and grey literature, are of key importance to the richness of the data-sharing environment.
13. Funding of generic tools supporting the administrative aspects of data sharing, such as access, authorization, as done through JISC and the e-Science Programme.



Data sharing strengthens and extends the role of data in the value chain. As the enhanced science and care it supports makes an impact, the funding agencies will in turn be encouraged to continue to provide sustained funding, which will in turn feed back into effective sharing and scientific advance. (See the diagram above.)

15.2 Models for the information flow

Data sharing involves a flow of information from producer to consumer as shown on the diagram on page 22. The following diagram is a re-drawing of that flow, simplified and indicating requirements at the various stages in the flow.



Examples of good practice have been noted throughout preceding sections, many of them cross-cutting and applicable to more than one point in the flow, and also to other issues not illustrated here. Funding implications are listed in section 15.1.

The following notes refer to the requirements of the three major players (producers, data resources, consumers) listed on the right of the diagram, observing general lessons and reference to sections where these are discussed further.

A Data planning: We have included this, even though it is not yet standard practice. Clearly producers need to know what policies are in place from their funders and host institutions to enable them to formulate these. They also need to know what course of action to take where policies and requirements may vary between co-funders or with the host institution. Producers

will require support, guidance and possibly training; there may be an allowance in grants for funding this activity. See sections 7, 11.

- B Data and metadata creation:** Section 4 emphasised the need for use of agreed standards for interoperability at a technical and at a semantic level. Metadata needs to be collected in this early phase, with support where available from tools. Data management planning can assist their identification. See sections 4, 6, 7.
- C Package and transfer:** Data needs somewhere to go for it be stored and made available for sharing; the producer needs to know this and it should be specified in the data management plan. The producer needs to be reassured that his rights will be respected in relation to the data in the future, and to know that any restrictions on access, use and confidentiality will remain in force. See sections 7, 8, 10, 12.
- D Ingest:** Data resources receiving data would prefer to know in advance it is coming so that they can plan capacity and advise the producer to submit information in suitable formats, give guidance on metadata needed and know what conditions of custody will be appropriate so as to ensure efficiency in subsequent data management and curation. See sections 7, 9.
- E Manage and curate information:** Management of a repository requires a measure of stability in funding. Support for staff, including training and appropriate career structures are required, depending on the levels of curation undertaken and the sophistication of the data provision made to consumers. See sections 9, 10, 11.
- F Discovery:** Data needs to be discoverable to be shared – seamless access is desirable, making it as easy for the consumer as possible. This does not just mean good user interfaces and searching tools, but also making it known the resource is there, it is easy to find, and put as few obstacles in the way as possible. Users will require support – a help desk on or-line help. Reliability and consistency are important too. See sections 5, 9, 10, 11.
- G Access and download data:** having obtained data, consumers should be able to assess it – to know what it is, what its status is, what rights and responsibilities are attached to it, and to be assured as to provenance, authenticity, and completeness. See sections 4, 8, 6, 12.
- H Re-use of data:** Use of standards and provision of sufficient and accurate documentation and other metadata are vital at this stage. In some circumstances restrictions on use will apply due to ethical considerations. Consumers may need support, either from the data resource or from the producer. See sections 4, 5, 6, 11.

15.3 Unresolved issues

We note here in conclusion that some issues still do not have complete solutions, and need further investigation and research:

- Preservation of digital data over the longer term (and its funding). The DCC, alongside other bodies is examining this question.
- Data selection (“appraisal”): What data do we need to keep; how can we judge its worth a priori?

- Provenance: How can we trust digital data when it is divorced from its original context? This is particularly acute where data are copied from one database to another, and the trail of provenance may be lacking, raising doubts about authenticity and veracity. In a clinical setting such data may be, literally, of vital importance.
- What to do with legacy data not in appropriate digital form? This includes the question of digitisation – one key informant confronts the question of whether or not to digitise research records which amount to some 95% of departmental research. Another is updating 55,000 records to be in line with controlled vocabulary recently agreed.

VOLUME 2 – Appendices

Table of contents

	Page
Appendix 1: Bibliography	A-2
Appendix 2: Glossary	A-10
Appendix 3: Case studies	A-19
Case study 1: Arts & Humanities Data Service	A-23
Case study 2: BADC / NERC DataGrid	A-29
Case study 3: BRIDGES	A-38
Case study 4: CLEF	A-47
Case study 5: Common Data Access	A-54
Case study 6: Ensembl	A-61
Case study 7: Genome Information Management System	A-67
Case study 8: Malaria (<i>Plasmodium falciparum</i>)	A-72
Case study 9: Nottingham Stock Arabidopsis Stock Centre	A-76
Case study 10: Proteomics Standards Initiative	A-83
Appendix 4: Further sharing models	A-97
Appendix 5: Supplementary materials	A-104
5.1 The data provenance issue	A-104
5.2 Standards and data sharing	A-109
5.3 Databases	A-113

Appendix 1

Bibliography

- [1] Academy of Medical Science, Science and innovation: working towards a ten-year investment framework - AMS response
- [2] Peter Aldhous, Prospect of data sharing gives brain mappers a headache, *Nature*, 2000
- [3] Sheila Anderson, Cataloguing and Documenting Historical Datasets: Proposals for the Next Millennium, Part III. Research Data Archives and Historical Data Bases, 1997
- [4] Peter Arzberger, Peter Schroeder, Anne Beaulieu, Geoff Bowker, Kathleen Casey, Leif Laaksonen, David Moorman, Paul Uhler, Paul Wouters. An international framework to promote access to data, *Science*, Vol. 303, 19 March 2004, 2004
- [5] Amit Bahl, Brian Brunk, Jonathan Crabtree, Martin J. Fraunholz, Bindu Gajria, Gregory R. Grant, PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data
- [6] Catherine A. Ball, Gavin Sherlock, Alvis Brazma, Funding high-throughput data sharing, *Nature Biotechnology*, Volume 22, Number 9, September 2004, 2004
- [7] MGED: Catherine Ball, Gavin Sherlock, Helen Parkinson, Philippe Rocca-Sera, Catherine Brooksbank, Helen C. Causton, et al., Editorial: An open letter to the scientific journals, *Bioinformatics*, Vol. 18. No. 11, 2002, p1409, 2002
- [8] Susan Bassion, The Clinical Data Interchange Standards Consortium Laboratory Model: Standardizing laboratory data interchange in clinical trials, *Drug Information Journal*, 2003
- [9] Rob Baxter, Denise Ecklund, Aileen Fleming, Alan Gray Brian Hills, Stephen Rutherford, Davy Virdee, Designing for Broadly Available Grid Data Access Services, *AHM*, 2003
- [10] Micha Bayer, Aileen Campbell, Davy Virdee, A GT3-based BLAST grid service for biomedical research, *Proceedings of the 2004 E-Science All Hands Meeting*, 2004
- [11] Kevin G.Becker, The sharing of cDNA microarray data, *Nature*, 2001
- [12] H.M. Berman, et al., The Protein Data Bank, *Nucleic Acids Res.*, 31, 23-27, 2003
- [13] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web: a new form of web content that is meaningful to computer will unleash a revolution of new possibilities, *Scientific American*, 284, 34-43, 2001
- [14] Andrew E. Berry, Malcolm J. Gardner, Gert-Jan Caspers, David S. Roos, Curation of the Plasmodium falciparum genome, *Trends in Parasitology*, Vol. 20, No. 12, December 2004, 2004
- [15] E. Birney, D. Andrews, P. Bevan, M. Caccamo, G. Cameron, Y. Chen, L. Clarke, G. Coates, T. Cox, et al., Ensembl 2004, *Nucleic Acids Research*, Vol. 32, Database issue, D468-470, 2004

- [16] Ewan Birney, Daniel T. Andrews, Paul Bevan, Mario Caccamo, Yuan Chen, Laura Clarke, Guy Coates, James Cuff, et al., An Overview of Ensembl, *Genome Research*, April 2004, pp 925-928, 2004
- [17] Christian Bourret, *Data concerns and challenges in health: networks, information systems and electronic records*, 2004
- [18] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, Ca.A. Ball, H.C. Causton, et al, Minimum information about a microarray experiment (MIAME) - towards standards for microarray data, *Nature Genet.* , 29, 365-371, 2001
- [19] Alvis Brazma, Editorial: On the importance of standardisation in the life sciences, *Bioinformatics*, vol. 17, no. 2, 2001, pp 113-114, 2001
- [20] Alvis Brazma, Alan Robinson, Graham Cameron, Michael Ashburner, One-stop shop for microarray data, *Nature* 403, 2000
- [21] Catherine Brooksbank, Graham Cameron, Janet Thornton, The European Bioinformatics Institute's data resources: towards systems biology, *Nucleic Acids Research*, Vol. 33, Database issue, 2005, 2005
- [22] British Standards SPSC Standards Policy and Strategy Committee, BS 0 - A standard for standards - Part 1: Requirements in the development of national standards and UK positions on international and European standards work - Draft for public comment, 2004
- [23] Steve Buckingham, Data's future shock, *Nature*, 2004
- [24] Daniel J. Carucci, Plasmodium post-genomics: an update, *Trends in Parasitology*, Vol. 20, No. 12, December 2004, 2004
- [25] Technical Committee CEN/TC 251 "Health Informatics", Health informatics - Electronic health record communication - Part 1 - reference model [re CEN TC251], 2004
- [26] Michele Clamp, Ensembl API Tutorial, Ensembl web site; revision: August 2004, 2004
- [27] Francis S. Collins, Eric D. Green, Alan E. Guttmacher, Mark S. Guyer, on behalf of the US National Human Genome Research Institute, A vision for the future of genomics research, *Nature*, Vol. 422, 24 April 2003, 2003
- [28] Brian M. Cooke and Ross L. Coppel, Blue skies or stormy weather: what lies ahead for malaria research?, *Trends in Parasitology*, Vol. 20, No. 12, December 2004, 2004
- [29] Mike Cornell, Norman W. Paton, Shengli Wu, Carole A. Goble, Crispin Miller, Paul Kirby; Karen Eilbeck, Andy Brass, Andrew Hayes, Stephen Oliver, GIMS - A data warehouse for storage and analysis of genome sequence and functional data
- [30] David J. Craigon, Nick James, John Okyere, Janet Higgins, Joan Jotham, Sean May, NASCArrays: a repository for microarray data generated by NASC's transcriptomics service, *Nucleic Acids Research*, vol 32, Database issue, D575-577, Oxford Unive, 2004
- [31] NERC DataGrid & EcoGRID projects - poster 2004
- [32] Committee on Responsibilities of Authorship in the Life Sciences, Sharing publication-related data and materials - responsibilities of authorship in the life sciences, 2003

- [33] Val Curwen, Eduardo Eyras, T. Daniel Andrews, Laura Clarke, Emmanuel Mongin, Emmanuel, Steven M.J. Searle, Michele Clamp, The Ensembl Automatic Gene Annotation System, *Genome Research*, 14: pp942-950
- [34] Rex Dalton, Young, worldly and unhelpful all miss out on data sharing, *Nature*, Vol. 404, 2 March 2000, 2000
- [35] Jo Dicks, Mary Anderson, Linda Cardle, Sam Cartinhour, Matthew Couchman, Guy Davenport, Jeremy Dickson, Mike Gale, David Marshall, Sean May, Hamish McWilliam, Andrew O'Malia, Helen Ougham, Martin Trick, Sean Walsh, Robbie Waugh, UK CropNet: a collection of databases and bioinformatics resources for crop plant genomics, *Nucleic Acids Research*, Vol. 28, No. 1, pp. 104-107, 2000
- [36] Peter Eckersley, Gary F. Egan, Shun-ichi Amari, Neuroscience data and tool sharing: a legal and policy framework for neuroinformatics, *Neuroinformatics*, Summer 2003, volume 1, issue 2, pp 149-166, 2003
- [37] Rasmus Fogh, John Ionides, Eldon Ulrich, Wayne Boucher, Wim Vranken, Jens P. Linge, Michael Habeck, Wolfgang Rieping, T.N. Bhat, John Westbrook, et al, The CCPN project: an interim report on a data model for the NMR community, *Nature structural biology*, volume 9, number 6, 2002
- [38] Michael Y. Galperin, The Molecular Biology Database Collection: 2005 update, *Nucleic Acids Research*, 2005, Vol. 33, Database issue, 2005
- [39] Daniel Gardner, and Gordon M. Shepherd, A gateway to the future of neuroinformatics, *Neuroinformatics*, 04/271-274, Humana Press, 2004
- [40] D. Gardner, A.W. Toga, G.A. Ascoli, et al., Towards effective and rewarding data sharing, *Neuroinformatics* 1, 289-295, 2003
- [41] Malcolm J. Gardner, Nell Hall, Eula Fung, Owen White, Matthew Berriman, Richard W. Hyman, Jane M. Carlton, Arnab Pain, Karen E. Nelson, Sharon Bowman, Ian T. Paulsen, Keith James, Jonathan A. Elsen, Kim Rutherford, et al, Genome sequence of the human malaria parasite *Plasmodium falciparum*, *Nature*, Vol. 419, pp.498-511, 2002
- [42] Christian Gieger, Hartwig Deneke, Juliane Fluck, The future of text mining in genome-based clinical research, *Drug Discovery Today: Biosilico*, Vol. 1, No. 3, 2003
- [43] Daniel H. Geschwind, Sharing gene expression data: an array of options, *Nature Reviews Neuroscience*, 2001
- [44] Carole Goble, Christopher Watson, Carole Goble discusses the impact of semantic technologies on the life sciences, *Drug Discovery Today: Biosilico*, Vol. 2, No. 1, 2004
- [45] John P. Helfrich, Knowledge management systems: coping with the data explosion, *Drug Discovery Today: Biosilico*, Vol. 2, No. 1
- [46] Christiane Hertz-Fowler, Chris S. Peacock, Valerie Wood, Martin Aslett, Arnaud Kerhornou, Paul Mooney, Adrian Tivey, Matthew Berriman, Neil Hall, et al., GeneDB: a resource for prokaryotic and eukaryotic organisms, *Nucleic Acids Research*, Vol. 32, Database issue, 2004
- [47] Tony Hey, Towards an e-Infrastructure for Research and Innovation: A Progress Report on e-Science, 2004
- [48] David E. Hill, Michael A. Brasch, Anthony A. del Campo, Lynn Doucette-Stamm, James I. Garrels, Judith Glaven, James L. Hartley, James R. Hudson Jr., Troy Moore, Marc Vidal, Academia-Industry Collaboration: an integral element for building "omic" resources, *Genome Research*, 14: pp 2010-2014, 2004

- [49] Jonathan Himlin, Shirley Williams, Rebecca Kush, HL7 and CDISC Renew Charter Agreement, Strengthen Relationship
- [50] HM Treasury, Department of Trade and Industry, Department for Education and Skills, Science and innovation: working towards a ten-year investment framework, 2004
- [51] HM Treasury, Department of Trade and Industry, Department for Education and Skills, Science and innovation: working towards a ten-year investment framework - Consultation document, 2004
- [52] Stephen L. Hoffman, G. Mani Subramanian, Frank H. Collins, Craig J. Venter, Plasmodium, human and Anopheles genomics and malaria, *Nature*, Vol. 415, pp. 702-709, 2002
- [53] Paul Horrocks, Sharon Bowman, Susan Kyes, , Andrew P. Waters, Alistair Craig, Entering the post-genomic era of malaria research, *Bulletin of the World Health Organization* 2000, 78, pp 1424-1437, 2000
- [54] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clarke, L., The Ensembl genome database project, *Nucleic Acids Research*, 2002, Vol. 30, No. 1, 38-41, 2002
- [55] T. Hubbard, D. Andrews, M. Caccamo, G. Cameron, Y. Chen, M. Clamp, L. Clarke, G. Coates, et al., Ensembl 2005, *Nucleic Acids Research*, Vol. 33, Database issue. D447-453, 2005
- [56] Insel, Thomas R. Insel, Nora D. Volkow, Ting-Kai Li, James F. Battey, Jr., and Story C. Landis, *Neuroscience Networks - Data-sharing in an Information Age*, 2003
- [57] Intra-governmental Group on Geographic Information, *The principles of Good Metadata management*, 2004
- [58] Jackson, Donald G., Healy, Matthew D., Davison, Daniel B., *Bioinformatics: not just for sequences anymore*, *Drug Discovery Today: Biosilico*, Vol. 1, No. 3, pp.103-104, 2003
- [59] H.V. Jagadish, University of Michigan; Oklen, Frank, Lawrence Berkeley National Laboratory, *Data management for the biosciences: Report of the NSF/NLM workshop on data management for molecular and cell biology*, Feb 2-3, 2003, 2003
- [60] Dongwon Jeong, , Doo-Kwon Baik, *A Practical Approach: Localization-Based Global Metadata Registry for Progressive Data Integration*, *Journal of Information & Knowledge Management*, Vol. 2, No. 4, 2003
- [61] Maggie Jones and Neil Beagrie, *Preservation management of digital materials*, 2001
- [62] D. Kalra, P. Singleton, D. Ingram, J. Milan, J. MacKay, D. Detmer, A. Rector, *Security and confidentiality approach for the Clinical E-Science Framework (CLEF)*, 2003
- [63] J.C. Kissinger, B.P. Brunk, J. Crabtree, M.J. Fraunholz, B. Gajria, A.J. Milgram, D.S. Pearson, J. Schug., A. Bahl, S.J. Diskin, et al., PlasmoDB: The Plasmodium genome resource., *Nature*, 419, 490–492. October 2002, 2002
- [64] Kerstin Kleese-van-Dam, et al, *NERC DataGrid Abstract - NDG web page*, NDG web site:
- [65] S.H. Koslow, *Should the neuroscience community make a paradigm shift to sharing primary data?*, *Nature Neuroscience*, Vol. 2, 863-261, 2000

- [66] Stephen H. Koslow, Michael D. Hirsch, Celebrating a decade of neuroscience databases: looking to the future of high-throughput data analysis, data integration, and discovery neuroscience, *Neuroinformatics*, Fall 2004, Volume 2, Issue 3, pp. 267-270, 2004
- [67] Stephen H. Koslow, Sharing primary data: a threat or asset to discovery?, *Nature Reviews Neuroscience*, Volume 3, April 2002, pp311-313
- [68] Rebecca D. Kush, C. David Hardison, How necessary are data standards?, *Scrip Magazine*, 2004
- [69] Philip Lord and Alison Macdonald, e-Science Curation Report: Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision, 2003
- [70] Christopher G. Love, Jacqueline Batley, Geraldine Lim, Andrew J. Robinson, David Savage, Daniel Singh, German C. Spangenberg, David Edwards, New computational tools for Brassica genome research, *Comparative and Functional Genomics*, Issue 5, pp.276-280, 2004
- [71] William W. Lowrance, Learning from experience: privacy and the secondary use of data in health research, 2002
- [72] Workshop, Database development for the *P. falciparum* genome project: report of a workshop 23 November 1998, Malaria.org/genome website, 1998
- [73] Maryann E. Martone, Amarnath Gupta, Mark H. Ellisman, e-Neuroscience: challenges and triumphs in integrating distributed data from molecules to human brains, *Nature Neuroscience*, Volume 7, Number 5, May 2004, pp467-472, 2004
- [74] Maryann E. Martone, Amarnath Gupta, Mark H. Ellisman, e-Neuroscience: challenges and triumphs in integrating distributed data from molecules to brains, *Nature Neuroscience*, Vol. 7, No. 5, 2004
- [75] Medical Research Council, Personal information in medical research, *MRC Ethics Series*, 2000
- [76] Medical Research Council et al., JDSS invitation to tender, version 2.2, 2003
- [77] Medical Research Council et al., JDSS invitation to tender, version 4.1, 2003
- [78] Maria Molina, Pak Yoong, Knowledge Sharing in a Co-Opetitive Environment: The Case of Business Clusters, *Journal of Information & Knowledge Management*, Vol. 2, No. 4, 2003
- [79] S.J. Nass and B.W. Stillman, Large-scale biomedical science: exploring strategies for future research, *National Academies Press*, 2003
- [80] National Biodiversity Network, *National Biodiversity Network 2002/2003 Annual Report*, 2003
- [81] ISO TC211: Geographic information
- [82] Editorial. Whose scans are they, anyway?, *Nature*, 2000
- [83] Editorial. A debate over fMRI data sharing, 2000
- [84] National Cancer Research Institute, *Strategic Framework for the Development of Cancer Research in the UK*, 2003
- [85] Natural Environment Research Council, *NERC Data Policy Handbook, Version 2.2 (December 2002)*, 2002

- [86] Eric K. Neumann, Eric Miller and John Wilbanks, What the semantic web could do for the life sciences, *Drug Discovery Today: Biosilico*, November 2004, Vol. 2, No. 6, 2004
- [87] Letter to Robert Hooke, 5 February 1676, *Correspondence of Isaac Newton*, ed. H.W. Turnbull, 1960
- [88] The Global Science Forum Neuroinformatics Working Group of the Organisation for Economic Co-operation and Development, Report on Neuroinformatics, from The Global Science Forum Neuroinformatics Working Group of the Organisation for Economic Co-operation and Development, 2002
- [89] Main points from OECD workshop on human genetic research databases - issues of privacy and security, 26-27 February 2004, 2004
- [90] OECD, OECD follow-up group on issues of access to publicly funded data - Interim Report, 2002
- [91] Kevin O'Neill, Ray Cramer, Maria Gutierrez, Kerstin Kleese van Dam, Siva Kondapalli, Susan Latham, Bryan Lawrence, Roy Lowry, Andrew Woolf, The Metadata Model of the NERC Datagrid
- [92] Norman W. Paton, Shakeel A. Khan, Conceptual modelling of genomic information, *Bioinformatics*, Oxford University Press, Vol. 16 no. 6, 2000, pp548-55, 2000
- [93] Report of the Public Health Sciences Working Group convened by the Wellcome Trust, The Wellcome Trust, 2004
- [94] PricewaterhouseCoopers, Global Data Management Survey 2004. Data Quality Management At the core of managing risk and improving corporate performance, 2004
- [95] Quackenbush, John; in: *Nature Biotechnology*, Volume 22, Number 5, May 2004, Data standards for 'omic' science, 2004
- [96] Alan Rector, Adel Taweel, Jeremy Rogers, David Ingram, Dipak Kalra, Robert Gaizauskas, Mark Hepple, Jo Milan, Richard Powers, Donia Scott, Peter Sing, *Joining up Health and Bioinformatics: e-Science meets e-Health*, 2004
- [97] J.H. Reichman, Paul F. Uhler, A contractually reconstructed research commons for scientific data in a highly protectionist intellectual property environment
- [98] Joel Richardson and Judith Blake, Data integration, Gene Ontology, and the Mouse, [web page - see AM's ontologies folder]
- [99] The Royal Society working group on intellectual property, *Keeping science open: the effects of intellectual property policy on the conduct of science*, 2003
- [100] Royal Statistical Society and UK Data Archive: Working Group on the Preservation and Sharing of statistical Material: Information for Data Producers, *Preserving and sharing statistical material*, 2002
- [101] Yucel Saygin, , Arnold Reisman, Yun Tong Wang, Value of Information Gained From Data Mining in the Context of Information Sharing, *IEEE Transactions on Engineering Management*, Vol. 51, No. 4, 2004
- [102] G.M. Shepherd, Supporting databases for neuroscience research, *Journal of Neuroscience* 22., 1497, 2002
- [103] G.M. Shepherd, et al., The Human Brain: neuroinformatics tools, for integrating, searching and modeling multidisciplinary neuroscience data, *Trends Neuroscience*, 460-468, 1998

- [104] Richard Sinnott, David Gilbert, David Berry, Ela Hunt, Malcolm Atkinson, Bridges: Security focused integration of distributed biomedical data, Proceedings of the 2003 E-Science All Hands Meeting, 2003
- [105] Richard Sinnott, Malcolm Atkinson, Micha Bayer, Dave Berry, Anna Dominiczak, Magnus Ferrier, David Gilbert, Neil Hanlon, Derek Houghton, Ela Hunt, David White, Grid services supporting the usage of secure federated, distributed biomedical data, Proceedings of the 2004 E-Science All Hands Meeting, 2004
- [106] Arne Stabenau, Graham McVicker, Craig Melsopp, Glenn Proctor, Michel Clamp, Ewan Birney, The Ensembl Core Software Libraries, Genome Research: 14: pp929-933 - Ensembl Special, 2004
- [107] James Stalker, Brian Gibbins, Patrick Meidl, James Smith, William Spooner, , Hans-Rudolf Hotz, Antony V. Cox, The Ensembl web site: mechanics of a genome browser ["advertisement" under], Genome Research, 951-955, 2004
- [108] Robert Stevens, Robin McEntire, Report on Sixth Annual Bio-Ontologies Meeting, Brisbane, 2003
- [109] Christian J. Stoeckert Jr., Helen C. Causton & Catherine A. Ball, Microarray databases: standards and ontologies, Nature Genetics Supplement, Vol. 32, December 2002, 2002
- [110] Christian J. Stoeckert, John Quackenbush Jr., Alvis Brazma, Catherine A. Ball, Minimum information about a functional genomics experiment: the state of microarray standards and their extension to other technologies, Drug Discovery Today, Vol. 3 No. 4, 2004
- [111] Nick Thomson, Artemis: the Goddess of the Hunt, Microbiology Today, vol. 30, February 2003, 2003
- [112] Arthur W. Toga, Neuroimage Databases: the good, the bad and the ugly, Nature Reviews Neuroscience, Volume 3, April 2002, pp 302-309, 2002
- [113] Performance and Innovation Unit (UK Government Dept.), Privacy and data sharing - the way forward for public services - a Performance and Innovation Unit Report, 2002
- [114] Open access and the public domain in digital data and information for science: Proceedings of an international symposium [held UNESCO, Paris in May 2003], 2004
- [115] D.C.Van Essen, Windows on the brain: the emerging role of atlases and databases in neuroscience, Current Opin. Neurobiol. 12, 574-579, 2002
- [116] John Darrell van Horn, Scott T. Grafton, Daniel Rockmore, Michael S. Gazzaniga, Sharing neuroimaging studies of human cognition, Nature Neuroscience, Volume 7, Number 5, May 2004, 2004
- [117] J.D. Van Horn, M.S. Gazzaniga, Databasing fMRI studies - towards a 'discovery science' of brain function, Nature Reviews Neuroscience, Volume 3, 2002
- [118] Ursula Weiss, Malaria: Foreword, Nature Insight Issue, February 2002, Nature. 415, 669. 7 February 2002, 2002
- [119] Wellcome Trust web site - malaria article
- [120] Wellcome Trust Sanger Institute malaria pages - 1
- [121] Wellcome Trust, MRC, Department of Health, UK Biobank ethics and governance framework, Version 1.0, 2003

- [122] The Wellcome Trust - report of a meeting organized by the Wellcome Trust, held 14-15 January 2003, Fort Lauderdale, USA, Sharing data from large-scale biological research projects: a system of tripartite responsibility, 2003
- [123] Matthew West, Some industrial experiences in the development and use of ontologies, 2004
- [124] Virginia A. de Wolf. In: Data Science Journal, Volume 2, Issues in accessing and sharing confidential survey and social science data, 2003
- [125] Andrew Woolf, Ray Cramer, Maria Gutierrez, Kerstin Kleese van Dam, Siva Kondapalli, Susan Latham, Bryan Lawrence, Roy Lowry, Kevin O'Neill, Data Virtualisation in the NERC Datagrid
- [126] Report of the Workshop on Best Practices in International Scientific Co-operation
- [127] Paul Wouters, Data Sharing Policies, OECD working group., 2002
- [128] ed. Paul Wouters, and Peter Schröder, Promise and practice in data sharing, 2003
- [129] Tadataka Yamada, Letter from SmithKline Beecham to House of Commons
- [130] Martin Yuille, Bernhard Korn, Troy Moore, , Andrew A. Farmer, John Carrino, Christa Prange, Yoshihide Hayashizaki, The responsibility to share: sharing the responsibility, Genome Research, 14: pp 2015-2019, 2004

N.B. The Bridges and PSI case studies (appendices 3.3 and 3.10 respectively) also include individual sets of references.

Appendix 2

Glossary: Abbreviations and technical terms

* at the beginning of a definition indicates that the term and its definition are specific to this report.

Term	Meaning
AceDB	AceDB is a genome database at the Sanger Institute designed specifically for handling bioinformatics data.
AHDS	Arts and Humanities Data Service
AHRC	Arts and Humanities Research Council (formerly the Arts and Humanities Research Board, AHRB)
AIIM	The Association for Information and Image Management
AMS	Academy of Medical Sciences
Annotation	A free text addition to a digital resource.
Anonymisation	To anonymise data means to destroy the link that exists between the data and the individual or object from whom it was taken, enabling identification of that individual or object.
API	Application Programming Interface (or Application Program Interface): The method prescribed by a computer system or programme for accessing another computer system or application.
Array Express	ArrayExpress is a public repository for microarray data, which is aimed at storing well annotated data in accordance with MGED (qv) recommendations.
ASCII	American Standard Code for Information Interchange
BADC	British Atmospheric Data Centre
BBSRC	Biotechnology and Biological Sciences Research Council
BGS	British Geological Survey
BioBank	UK Biobank is a long-term national project to build the world's largest information resource for medical researchers. It will follow the health of 500,000 volunteers aged 40-69 in the UK for up to 30 years.
BLAST	Basic Local Alignment Search Tool. A method for rapid searching of nucleotide and protein databases.
BRC	Bioinformatics Research Centre, University of Glasgow
BRIDGES	Biomedical Research Informatics Delivered by Grid Enabled Services
CAD	Computer-Aided Design
CADRE	Central Aspergillus Data Repository. A repository to manage and curate <i>A. fumigatus</i> genomic data. In addition, since the species within the genus <i>Aspergillus</i> have recently evolved from each other, it enables comparative studies by gathering other <i>Aspergilli</i> genomic data into a central data repository.
CCLRC	Council for the Central Laboratory of the Research Councils
CDA	Common Data Access

Term	Meaning
CDISC	Clinical Data Interchange Standards Consortium. CDISC is developing a global, platform-independent data standard to enable information system interoperability in medical research and related areas of healthcare. See http://www.cdisc.org/
CEN	European Committee for Standardization
CERN	Conseil Européen pour la Recherche Nucléaire (European Council for Nuclear Research), now usually understood as the European Organization for Nuclear Research.
CLEF	Clinical eScience Framework
CODATA	Committee on Data for Science and Technology
Community resource	A data resource which stores, maintains, curates and makes available data sharing resources for a community of users.
Compression	A set of techniques used to reduce the size of a digital file. Some methods are lossless (no information is lost on decompression after compression), others can lose some information.
Consumer	Individual or organisation which receives data (sometimes referred to as a customer, user or recipient).
Co-ode	Collaborative Open Ontology Development Environment project, research team at the Medical Informatics Group at the University of Manchester.
COREC	Central Office for Research Ethics Committees
Creative Commons	A non-profit organization devoted to expanding the range of creative work available for others to legally build upon and share. See http://creativecommons.org/
Curation	The activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and with other published materials (definition from e-Science Curation report; see also discussion of definition of digital curation at the DCC web site, www.dcc.ac.uk)
Data	In this study, data in digital form, whether raw data newly gathered (by whatever method, including simulation and modelling), and data derived from that by processes of consolidation, selection, calculation, and statistical analysis, etc.
Data conversion	The process of changing data from one format to another (e.g. from, say, Microsoft Word to PDF)
Data farms	*Data resources which offer datasets in the form of whole, discrete files for downloading.
Data resources	*Organisations which store and/or make data available data to producers. They receive data from consumers; they may also create their own, additional, data to add to consumers' data. A data producer can also be a data resource.
DDBJ	DNA Databank of Japan
Depersonalisation	Removing information which might help identify an individual, such as addresses, telephone numbers, profession. See also anonymisation.
DEST	Department of Education, Science and Training

Term	Meaning
Discovery	*The process, by consumers (q.v.), of finding information or data from data resources.
DOI	Digital Object Identifiers. One form of persistent digital object identifier.
DTI	Department of Trade and Industry
eBank UK	eBank is a project to explore the potential for integrating research data sets into digital libraries by using common technologies such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Development outcomes from the project include the distributed information architecture, requirements for common ontologies, data models, metadata schema, open linking technologies.
EBI	European Bioinformatics Institute
EGTDC	Environmental Genomics Thematic Programme Data Centre. A NERC initiative. In March 2005, the EGTDC formally changed its name to the NERC Environmental Bioinformatics Centre (NEBC).
e-Infrastructure	The networks, software elements, computers and other components on which sharing takes place. These are provided, variously, on local, national and international scales.
EMBL	European Molecular Biology Laboratory
Ensembl	Ensembl is a joint project between the EMBL-EBI and the Wellcome Trust Sanger Institute that aims to develop a system that maintains automatic annotation of large eukaryotic genomes (q.v. eukaryotes).
EnvGen	NERC's Environmental Genomics research programme
EpoDB	Erythropoiesis database is a database of genes that relate to vertebrate red blood cells.
ESRC	Economic and Social Research Council
Eukaryotes	Organisms with "good nuclei" in their cells, i.e. animals, plants and fungi, as opposed to prokaryotic cells of bacteria and blue-green algae)
FASTA	A computer program based on the method of W. Pearson and D. Lipman to search for similarities between one DNA or protein sequence and any group or library of sequences. FASTA is pronounced "FAST-Aye", and stands for "FAST-All", because it works with any alphabet, an extension of "FAST-P" (protein) and "FAST-N" (nucleotide) alignment.
Firewall	A system designed to prevent unauthorized access to or from a private network.
FITS	Flexible Image Transfer System. A commonly used file format in astronomy
Flat files	Files with no internal hierarchical structure. Such files can be read as a single stream of information, like a plain text file.
fMRIDC	Functional MRI Data Centre. http://www.fmridc.org/f/fmridc (q.v. for MRI)
FTP	File Transfer Protocol. The Internet protocol that permits you to transfer files between your system and another system.
GBIF	Global Biodiversity Information Facility.
Gbits	Gigabits: one billion bits. The term is commonly used for measuring the amount of data that is transferred in a second between two telecommunication points. Gigabits per second is usually shortened to Gbps.
Gigabyte	2 to the 30th power (1,073,741,824) bytes. One gigabyte is equal to 1,024 megabytes.

Term	Meaning
GIMS	Genome Information Management System
GIS	Standard ISO 19115:2003 defines the schema required for describing Geographic Information and Services.
Globus Alliance	A community of organizations and individuals developing technologies behind the "Grid," which let people share computing power, databases, instruments, and other on-line tools securely across corporate, institutional, and geographic boundaries.
GNU	UNIX-compatible operating system developed by the Free Software Foundation. From which software distributed under the GNU Public License (GPL).
GriPhyN	Grid Physics Network. US-based collaborative project. See: http://www.griphyn.org/
GTAC	Gene Therapy Advisory Committee
GUS	Genomics Unified Schema, developed by the University of Pennsylvania
HapMap	The International HapMap Project is developing a haplotype map of the human genome, the HapMap, which will describe the common patterns of human DNA sequence variation.
Heap	A rapid access temporary storage area storage area used by computer programs.
HFEA	Human Fertilisation and Embryology Authority
HTA	Human Tissue Authority
HTML	HyperText Mark-up Language
HTTP	HyperText Transport Protocol. Used to transfer HTML documents across the World Wide Web.
HUPO	Human Proteome Organization. Engaged in scientific and educational activities to encourage the spread of proteomics technologies and to disseminate knowledge pertaining to the human proteome and that of model organisms.
ICT	Information and Communications Technology
IGGI	Intra-governmental Group on Geographic Information
Information	In this report, data viewed in context, and endowed with meaning by a data user.
Ingest	An OAIS (qv) term for the process of adding data to an archive.
ISO	International Organization for Standardization
ISO 15926	A standard for a conceptual data model for computer representation of technical information about process plants.
IT	Information Technology
JCSR	JISC Committee for the Support of Research
JISC	Joint Information Systems Committee
JPEG	Joint Photographic Expert Group. A graphics standard.
MAGE-ML	MicroArray Gene Expression Mark-up Language. Describes how the content and form of microarray experiment metadata are to be expressed in XML.
MAGE-OM	MicroArray Gene Expression Object Model. A UML model to describe microarray experiments.

Term	Meaning
maxdLoad2	A standards-compliant database development which supports the export of microarray data in the MAGE-ML format, for submission to the Array Express database.
Metadata	This is usually defined as data which describes or qualifies other data. Here it includes descriptions (of all kinds), annotation, documentation and commentary.
Methods	Some data may make no sense when shared unless the methods and workflows used to process and view it are also shared; usually these will be delivered to the recipient in the sharing process either as metadata or encoded/implied in software tools
MGED	Microarray Gene Expression Database society
MHRA	1. Modern Humanities Research Association 2. Medicines and Healthcare products Regulatory Agency
MIAME	Minimum Information about A Microarray Experiment. A data content standard; it does not prescribe a format for the metadata it specifies
MIAPE	Minimum Information about a Proteomics Experiment A statement of what should be recorded about microarray experiments. See MAGE-OM/ML
Microarray	A method for profiling gene and protein expression in cells and tissues
MIPS	Munich Information Centre for Protein Sequences
mmCIF	macromolecular Crystallographic Information File. CIF is a subset of STAR (Self-defining Text Archive and Retrieval format). The CIF format is suitable for archiving, in any order, all types of text and numerical data.
MRC	Medical Research Council
MREC	Multi-centre Research Ethics Committee. See also COREC and REC.
MRI	Magnetic Resonance Imaging
myGrid	myGrid is a toolkit of core components for forming, executing, managing and sharing discovery experiments. See http://www.mygrid.org.uk/
NASA	National Aeronautics and Space Administration
NASC	Nottingham Arabidopsis Stock Centre
NCBI	National Center for Biotechnology Information
NCII	National Cancer Informatics Initiative (from the NCRI)
NCRI	National Cancer Research Institute
NDG	NERC DataGrid
NEBC	NERC Environmental Bioinformatics Centre (Formerly the EGTDC, qv)
NERC	Natural Environment Research Council
NHDA	National Hydrocarbon Data Archive
NHS	National Health Service
NIH	National Institutes of Health
NPSA	National Patient Safety Agency
NTRAC	National Translational Cancer Research Network
OAIS	Open Archival Information System, a reference model for archive systems, ISO standard ISO14721

Term	Meaning
OECD	Organisation for Economic Cooperation and Development
OGSA-DAI	The OGSA-DAI project is concerned with constructing middleware to assist with access and integration of data from separate data sources via the grid.
OGSA-DAIT	Continuation of the original OGSA-DAI project.
OMG	Object Management Group, a consortium formed to improve the interoperability of software systems
OMII	Open Middleware Infrastructure Institute
OMIM	Online Mendelian Inheritance in Man. A catalogue of human genes and genetic disorders authored and edited by Dr. Victor A. McKusick and colleagues at Johns Hopkins and elsewhere, and provided through NCBI (q.v.). The database contains information on disease phenotypes and genes, including extensive descriptions, gene names, inheritance patterns, map locations and gene polymorphisms
Ontology	<ol style="list-style-type: none"> 1. Concepts, and their relationships expressed in a formal language 2. A tool to express and exploit an ontology per definition 1.
openEHR	Project to publish formal specification of requirements for representing and communicating electronic health record information. See: http://www.openehr.org/index.html
Parse	To break down a sequence of letters or numbers into meaningful parts based on their location in the character sequence. To break character string or file into its elemental components for the purpose of interpretation.
Parse	<ol style="list-style-type: none"> 1) To analyze a sentence or phrase, stating its part of speech, form, syntactical relationships. 2) In a digital context, to analyze and break down words into functional units that can be converted into machine language.
PDB	Protein Data Bank
PEDRo	A database for storing, searching and disseminating experimental proteomics data (Not to be confused with the Physiotherapy Evidence Database, which also uses the same acronym.)
PERL	Practical Extraction and Report Language. Perl is a programming is a computer language, which makes it easy to build and test simple programs.
PERMIS	Privilege and Role Management Infrastructure Standards Validation. Open-source software from the NSF's Middleware Initiative for access controls to systems and data.
Petabyte	2 to the 50th power (1,125,899,906,842,624) bytes. A Petabyte is equal to 1,024 Terabytes.
PI	Principal investigator
PIAG	Patient Information Advisory Group
Post-genomic	After the determination of a genome
PPARC	Particle Physics and Astronomy Research Council
Producer	*of what is to be shared (and who is often its creator)
Pseudonymisation	The replacement of all data which can identify the individual to whom it relates with information which, for that individual, bears the same relation to each other as in the original information.

Term	Meaning
PSI	Proteomics Standards Initiative
QTL	Quantative Trait Loci
RAE	Research Assessment Exercise. In the UK higher education sector, to assess the quality of UK research and to inform the selective distribution of public funds for research by the UK higher education funding bodies.
RAL	Rutherford Appleton Research Laboratory
RDBMS	Relational Database Management System. A type of DBMS (database management system) in which the database is organized and accessed according to the relationships between data values held in tables. Also called RDMS.
RDF	Resource Description Framework. A general framework for how to describe any Internet resource.
RDL	1) Reference Data Library – part of ISO 15926 2) Report Definition Language, a XML-based language created by Microsoft with the goal of promoting interoperability of commercial reporting products.
REC	Research Ethics Committee
Repository	A place or facility where things or data can be stored, possibly providing an index to its contents
SAM	Scientific Annotation Middleware. Collaborative project in the USA, under the Scientific Discovery through Advanced Computing (SciDAC) initiative.
SCAG	Security and Confidentiality Advisory Group. A group was established in 1996 to govern access to three national NHS databases which hold patient information: the Hospital Episode Statistics database; the NHS-Wide Clearing Service database; and the NHS Strategic Tracing Service.
Schema	The structure of a database system, document or other data construct, described in a formal language.
SDSS	Sloan Digital Sky Survey
Sharing	*For this study, we take data sharing to mean the re-use of data, in whatever way, and wherever it is stored.
Sharing - asynchronous	*When shared with an interval between use and re-use – perhaps amounting to decades or centuries (as when archived).
Sharing - consecutive	*When data is passed from one person or body to another sequentially (this form of data sharing is common in the pharmaceuticals and oil industries, and so too in patient care).
Sharing - in situ	*Where data is exploited by a consumer (qv) on-line, the data staying at the data resource,
Sharing - parallel	*Simultaneous sharing by many people
Sharing - pull	* Consumers seeking and downloading information to themselves.
Sharing - push	*Consumers being sent information from the producer or data resource, the data resource initiating the transaction.
Sharing - remote	Where the consumer takes a copy or extract of data from a resource, downloads it and uses it locally.
Sharing - synchronous	When it is more or less contemporaneous with data creation

Term	Meaning
Sharing: large-scale	<p>*Means in this context that one or more of the following hold:</p> <ul style="list-style-type: none"> ■ volumes are large (gigabytes and beyond); ■ the number files or records are very large (e.g. in millions); ■ the number of transactions which take place is large (e.g. hit rates on a web site); or ■ the number and/or diversity of participants is large
SHWFGF	Sir Henry Wellcome Functional Genomics Facility, at the University of Glasgow.
SkyServer	A website presenting data from the Sloan Digital Sky Survey (SDSS), a project to make a map of a large part of the universe available to all.
SNP	Single nucleotide polymorphisms are common DNA sequence variations among individuals.
SNP Consortium	The SNP Consortium (TSC) is a public/private collaboration that has discovered and characterized nearly 1.8 million SNPs.
SOAP	Simple Object Access Protocol. A protocol originally designed by Microsoft and IBM for exchanging messages between computer software.
SQL	Structured Query Language
SRB	Storage Resource Broker. A data storage management system from the University of California Supercomputer Center at San Diego.
SRS	<p>1) SRS is a free searching resource for molecular biology data. It has two functions: a data retrieval system for more than 140 databases incorporated into the EBI's SRS server and also as a data analysis applications server.</p> <p>2) Synchrotron Radiation Source. A facility for the exploitation of Synchrotron Radiation (SR) for fundamental and applied research at Daresbury, UK.</p>
STEP	Standard for the Exchange of Product Model Data. An ISO standard (ISO 10303 et seq) that describes how to represent and exchange digital product information.
Support	Assistance given to users of a resource by those with specialist knowledge of that resource.
TAIR	The Arabidopsis Information Resource
Terabyte	2 to the 40th power (1,099,511,627,776) bytes. A Terabyte is equivalent to a 1,024 Gigabytes
TIFF	Tagged Image File Format. A standard for graphics files.
TIGR	The Institute for Genome Research
Tools	Software for data discovery, visualisation, presentation, processing in the context of re-use, storage, archiving and preservation, and the management of sharing processes
UKOLN	A UK centre of expertise in digital information management, providing advice and services to the library, information, education and cultural heritage communities. The acronym originally stood for the United Kingdom Office of Library Networking.
UML	Unified Modelling Language, a standard notation for representing software systems in a platform-independent manner.

Term	Meaning
UniProt	Universal Protein Resource. A comprehensive catalogue of information on proteins, protein sequences and function hosted by the EBI. See http://www.ebi.uniprot.org/index.shtml
VDL	Virtual Data Language. (See GriPhyN, q.v.)
VOTES	Virtual Organisations for Trials and Epidemiological Studies. A collaborative project involving the universities of Oxford, Glasgow, Leicester, Manchester, Nottingham and Salford; the National e-Science Centre, the London e-Science Centre and Imperial College London. The project is funded by the Medical Research Council.
W3C	World-Wide Web Consortium
WSDL	Web Services Definition Language
X.509	A widely used standard for defining digital certificates, which provide a public key infrastructure for secure communications.
XML	eXtensible Mark-up Language
Xpath	XML Path Language. XPath is the result of an W3C effort to provide a common syntax and semantics for functionality XML pointers (cross references).
Xquery	XML Query Language. a query language called XQuery, which is designed to be broadly applicable across many types of XML data sources.
XSL	Extensible Stylesheet Language. A language for transforming XML documents into other document formats.
XSLT	Extensible Stylesheet Language Transformations. An XML standard from W3C defining the syntax and semantics of a language for transforming XML documents into other XML documents. See also XSL.

Appendix 3

Case studies

The report looked at ten major case studies. The case studies were selected to cover the domains and data types shown in Box A-1, including both community resources and principal investigator-led activities. They span projects of different ages and a range of points along the data-sharing/data management flow, from different perspectives, and illustrating different organizational models.

Box A-1: Domains and data types covered by the study

<p>Domains:</p> <ul style="list-style-type: none"> ■ Human, animal, plant sciences, primarily focusing on: <ul style="list-style-type: none"> - Post-genomic research - Population research (health, disease, biodiversity) ■ In addition, selected exemplars from the environmental domain and comparators from outside the life sciences. 	<p>Data types:</p> <ul style="list-style-type: none"> ■ Genome sequence ■ Micro arrays ■ Biological structures ■ Laboratory and clinical data (including imaging) ■ Field and population
---	--

After considering a large number of candidates a short-list was drawn up using the criteria noted above, and the final list provided to the sponsors for agreement.

The major case studies

In alphabetical order, the eight main case studies – and some of the features for their inclusion – followed by two comparator studies:

The first case study (BRIDGES) was used as a pilot, and proved to be very instructive not only from a methodological perspective but also from an information view.

- **BRIDGES** (Bio-Medical Research Informatics Delivered by Grid-Enabled Services): a supporting project for a large biomedical research project, Cardiovascular Functional Genomics (CFG) investigating the genetic causes of hypertension (one of the major causes of mortality and disease in the western world). The BRIDGES project aims to develop an infrastructure (using the GRID) which facilitates improved CFG life science research through integrating and/or federating a broad range and class of biomedical data sources, and providing a suite of tools for exploiting this data, and as such tailored to investigating functional genomics research. The case study represents a concrete exploration of what can be achieved now in terms of access to and usage of a variety of life-science data resources.
- **BADC**: The British Atmospheric Data Centre (BADC)/Nerc DataGrid (NDG): the BADC is one of NERC's three largest thematic data centres, holding UK atmospheric data generated by NERC for researchers and the geo-communities. As such it is custodian of unique, non-reproducible observational data, a substantial proportion from sensors and satellites. It has many years' experience as a data centre providing services to its users, working on curation and preservation of its holdings. It plays a large role in the Nerc DataGrid project which, like BRIDGES, is an opportunity to consider the issues which arise in collaborative, federated data-sharing environments, giving an opportunity to explore (for example) many years' experience in standards development, metadata, retrieval issues.
- **CLEF**: Clinical e-Science Framework: this project is developing an "end-to-end" information framework to link health-care delivery, clinical and post-genomic research. Its goal for research is a pseudonymized repository of chronicles of patient records, to aid patient care, aggregated for post-genomic research. This project therefore provides an opportunity to consider clinical data in research and health care contexts, and key issues such as respect of consent and confidentiality.
- **Ensembl**: a comprehensive and integrated source of annotation of large genome sequences, tools, maintained by the Wellcome Trust Sanger Institute and the EBI. In 2004 the number of genomes rose from nine to 16 and now stands at [18], providing wide and deep comparative analysis for users, the web interfaces allowing direct comparison of the annotation of different genomes. The Ensembl software itself is increasingly adopted in third-party projects around the globe.
- **GIMS**: The Genome Information Management System (GIMS): An example of a PI-led resource springing from its unit's previous work on a community resource (yeast) and being taken on to a next generation of work (fungi). It is also a user of other resources, providing a multiple perspective. It is also interesting from a database and tools point of view, taking in data from disparate sources and combining them in a data warehouse, applying an information model, and providing "canned" queries for more powerful analysis.
- **Malaria DB**: The *Plasmodium falciparum* malaria genome: of interest in particular for its genesis – the assertion of a community resource open to all, its user-focused database development – as well as an example of a curated database (at the Sanger Institute) within a wider resource.

- **NASC:** The Nottingham Arabidopsis Stock Centre (NASC): In the plant sciences: provides microarray services, collects and curates data for customers which are also held in the NASC repository; provides tools (including tools based on Ensembl).
- **PSI:** The Proteomics Standards Initiative: Set up by HUPO, the Human Proteome Organization, in 2002 to address the data-sharing challenges that arise in proteomics: data sharing is hindered by the lack of major central databases for experimental protein data and no widely accepted standard formats for exchanging data between groups. An opportunity to look closely at standards development, the role of industry, and forward challenges, including co-ordination with other “omics”.

The comparator case studies

These case studies chosen from outside the life sciences, one in the academic sector and the other in industry.

- **Comparator: AHDS:** The Arts and Humanities Data Service (AHDS) – The AHDS has one of the longest histories as a publicly funded data service in the UK, with considerable experience in management and organization of a data service, as well as data management, preservation, and curation. It handles a wide range of data formats, from simple to complex. A multi-community resource.
- **Comparator: CDA:** Common Data Access Limited (also known as CDA Limited): Serving the corporate sector and also using data and services from NERC, CDA is a not-for-profit subsidiary of the UK Offshore Oil Operators Association (with DTI backing); it provides a data service for oil well/drilling data and seismic data relating to the UK Continental Shelf area, to fee-paying members (mostly companies) and also provides public access to data via the web; the data is provided by members and by the British Geological Survey (a NERC unit).

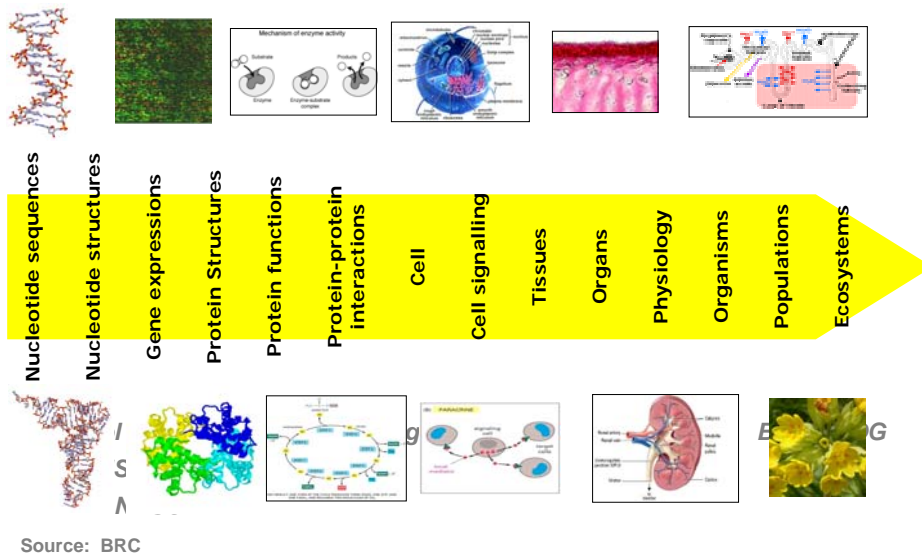
The supplementary case studies

These case studies listed above led us to look in depth at other resources which might themselves have been the subject of a case study – initiatives such as the Gene Ontology, UniProt, MGED and MIAME, The International SNP Consortium, and other Grid initiatives such as MyGrid.

The study team also conducted several mini case studies, through face-to-face interviews and desk research: GBIF, OpenEHR, NTRAC, NCRI, CDISC, brain neuro-imaging; STEP and ISO 10303/15926; Open Geospatial Standard.

The following diagram indicates where the case studies fall in the spectrum of the life sciences, from the genome to the environment. This is just a rough indication, as many of them spanned a range of biological entities. (The CDA and AHDS do not of course fit into this spectrum.)

Life sciences: a complex landscape



Detailed case study reports

The following section presents detailed reports, in the form of short essays, on each major case study. They are presented in alphabetical order. Each case study report is prefaced with a summary of key data.

Appendix 3.1 Arts and Humanities Data Service

Key data

Name:	Arts and Humanities Data Service
Acronym:	AHDS
Established:	1996
Brief description:	<p>The Arts and Humanities Data Service (AHDS) is the UK's national service aiding the discovery, creation and preservation of digital resources in and for research, teaching and learning in the arts and humanities. The Service has a Managing Executive at Kings College London, and this supports five subject centres, as follows:</p> <p>AHDS Archaeology (Archaeology Data Service at the University of York, http:// ahds.ac.uk/archaeology/)</p> <p>AHDS History (University of Essex, http:// ahds.ac.uk/history/)</p> <p>AHDS Visual Arts (University College for the Creative Arts, http:// ahds.ac.uk/visualarts/)</p> <p>AHDS Literature, Languages and Linguistics (Oxford Text Archive at the University of Oxford, http://.ahds.ac.uk/litlangling/)</p> <p>AHDS Performing Arts (University of Glasgow, http://ahds.ac.uk/performingarts/)</p>
Contact:	<p>Ms Sheila Anderson (Director) Kings College, London</p> <p>http://ahds.ac.uk</p>
Funding:	£1M pa (2004), from JISC, and AHRC. The Executive and the five subject Centres are also supported by their host institutions.
Staffing:	12 staff at Kings College, London, some four staff at each of the satellites.
Interview data:	Sheila Anderson, Hamish James, Mathew Woollard (AHDS History) were interviewed by Alison Macdonald and Philip Lord, 15 November 2004.

AHDS Case study report

The Arts and Humanities Data Service (AHDS) was established in 1996 to collect, preserve and encourage the reuse of digital resources created during scholarly research from the arts and humanities. In late 2004 the AHDS was responsible for the preservation of over 3,000 digital resources and holds a wide range of data types, from plain text and image files to datasets (600 spreadsheets, databases, statistical data files) and digital recordings (audio or audio/video), as well as more complex resources such as Web sites and GIS (Geographical Information System) data.

The AHDS is organised as a distributed service, consisting of a managing executive, hosted by King's College London, and five subject centres (see above).

Each AHDS Centre takes responsibility for advising on and “ingesting” digital resources in its subject area. In the past, each centre made separate arrangements for the storage, preservation and delivery of its collections. One of the unanticipated results of this was that each Centre adopted a different approach. AHDS History and AHDS Archaeology emphasised long-term preservation, while the other three Centres focused more on providing online access to their holdings. Now, as part of a move to a more centralised organisational model, the AHDS is developing an archival digital repository that will be used by all five AHDS Centres to preserve their collections, and will serve as the back-end to a new range of common access methods provided to our users.

The new repository is to be built around a data store of approximately 12TB (10TB tape library, 2TB disk array) with offsite replication. The current collection size is approximately 1TB, with 2-3TB due to arrive in the very near future. They are planning to use the Storage Resource Broker (SRB) from the San Diego Supercomputer Center (SDSC) to manage staff access to these data stores. Having considered products like FEDORA and DSpace, the AHDS is using more generic Java and XML (Cocoon) technologies to develop online tools for delivering data to its users. Generally, these will replace existing tools developed at different AHDS Centres, which include tools for downloading files, viewing images, map based searching, dataset query and so on, though in some specific cases the new tools will work in parallel with the existing tools. They are using METS to package various metadata schemas (DDI, TEI, and internal schemas for resource discovery and preservation), and will provide a range of schemas via OAI.

Their funders are chiefly the Arts and Humanities Research Council (AHRC) and the Joint Information Systems Committee (JISC). Support is also provided by the six host institutions. AHDS central funding is currently funded at ca. £1M per annum.

The AHDS History group is a model for the other AHDS centres: it has four professional staff, a manager (Mathew Woollard), and two people (students) doing more every-day data processing work. The four professionals deal respectively with acquisitions, metadata, ingest of data, technical issues. These same roles tend to be replicated across all the AHDS units.

It was noted there are no accepted levels of professionalism or qualifications in the data curation area, but they are needed.

Regarding deposit of data in the AHDS repositories, if work is funded by the AHRC then funding recipients must offer any data produced to the AHDS. The AHDS evaluates offers

and has the right to refuse to take material. In some cases, the AHDS will waive the requirement for data to be offered, for instance, if data has to be offered to another archive (such as The National Archive), or in light of specific individual circumstances. Generally, the AHDS takes over data when a resource becomes “fixed”, though it also takes updates of data, if the deposit is being updated.

On a practical note: they always keep the originals that they receive: By doing this if there are IP or copyright problems one can merely give this back, whilst any information added afterwards contains the intellectual property from the repository and curatorial processes.

Regarding the quality of documentation (metadata) they received, our informants noted people lack time to do it properly and initially they do not see the point of having to supply it. Metadata collection therefore has been a problem. They have forms for submitters, but obtaining accuracy is difficult. The AHDS teams can add preservation metadata but not resource discovery metadata which pertains to the content of the resource. A solution is enforcing early provision of metadata, where the researchers must assign metadata values early in the data lifecycle. The AHDS points out to researchers that at some time in the future the submitter himself or herself might need to use the data too, and might need to find and use the information. They noted that people tend to be “slow but keen” to submit their data, and indeed there was a pride expressed in having data “that deserves to be kept”.

One of the functions of the five dispersed subject centres and the central Executive is to provide users with advice. This is backed up by a very good website that is clear, easy to use and provides much information. Perhaps as one measure of their success they now receive many enquiries about software as people start up a project – usually asking “what software should I use for my project?” While making suggestions on software tools to use (or indeed to avoid), they seek to give more holistic advice about the process of creating digital material e.g. on data creation, metadata, sustainability, delivery methods. They take a life-cycle approach when giving advice – advising on data creation through to preservation. The AHDS runs workshops and other events, issues a newsletter twice a year, and maintains several mailing lists. The AHDS also produces technical papers. It regards their production as an ongoing process; the papers report on and discuss current issues, and they are made available through the AHDS web site.

They have done no usability testing on databases, but have done some testing on web-based front ends.

They noted the web has broadened the audience for data, but too many users do not have the skills or knowledge to use it meaningfully. People expect interpretative layers to be provided for them. They wondered whether we need to introduce more formality into the sharing process again?

The AHDS recommends standards, but does not mandate them; rather they encourage good practice, and encourage the use of some technical standards such as: TEI (Text Encoding Initiative). Launched in 1987, the TEI is an international and interdisciplinary XML-based standard that helps libraries, museums, publishers, and individual scholars represent all kinds of literary and linguistic texts for online research and teaching, using an encoding scheme that is maximally expressive and minimally obsolescent.

- The AHDS (rather than its users directly) uses DDI - The Data Documentation Initiative, an effort to establish an international XML-based standard for the content, presentation, transport, and preservation of documentation for datasets in the social and behavioural sciences.
- The AHDS uses METS – Metadata Encoding and Transmission standard. The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium.

In archaeology there are standards. For example, Sites and Monuments Records (SMR), are kept to describe physical objects – a metadata record of the object. The need for SMRs was driven by the statutory needs of UK local authorities, but it is still a rather loose standard which is in the process of being given more uniformity.

An interviewee noted that unfortunately people tend to conflate standards for metadata and standards for data, pointing out that one man's data is another man's metadata. As an example, The National Archives (TNA) has one million scanned wills and a catalogue of these. The catalogue therefore contains metadata to the scanned documents; to some researchers, however, this catalogue is itself data. Often there are somewhat arbitrary decisions about what is put in data or in metadata.

Regarding low-level technical standards – they do not present much of a problem. They do receive some odd/old media (media is a separate issue from formats), and they do have some “problematic” formats – for example virtual reality presentations. There is also an increasing move for people to mount interactive web sites, and these might be a problem to keep going in the future. They noted that there is not much thought given to interoperability and preservation issues when digital objects are created. Courseware could be difficult in the future, but so far they do not get much of it (however see the JISC Online Repository for Learning and Teaching Materials (JORUM) project).

A specific technical point was raised: numerical data can be made to look useable, when in fact it is not. For example data could be put into single Oracle table, and this taken and preserved. Other people could then take this data and would be able to use it. But could it be used meaningfully when context may be missing?

In summary, technical/lower levels for file formats and have practical solutions – though these might be time consuming or costly to apply. It is higher-level issues that are more difficult to resolve. Furthermore, they noted “the more useful the data - the less easy it is to standardise”!

When asked if they saw any impact due to the Freedom of Information act they said “yes”, in sense that in the universities it is having a job creation effect!

There is very variable use made of the AHDS resources. Users can download data, but thereafter there is no tracking of data use. Small downloads cause no impact on the storage/server resource, but if there are large volumes then there may be an impact, particularly if old data has to be converted, for example, to SPSS formats.

We discussed data sharing in general: An interviewee noted that the term “sharing” is an interesting word in this context: It could be interpreted as people using a resource

simultaneously and collaboratively, or alternatively it may mean re-use, when people act individually. An example of the former sharing use in the AHDS's context might be when a performance, say, of a play, is being put together. Hamish James noted that one needed to know something exists before it can be shared. There is also a question of how much to push data to people, or vice versa, people "pushed" to data.

The AHDS people noted that despite their long experience there was very little interest in what they were doing from the science-based curation centres, except from the UKDA at Essex University and from the MRC.

There is a lot of cooperation between the arts (and artists) and engineers, for example. This example highlights some problems raised by cross-domain activity – for instance funding structures do not generally foster such cooperation: There is the question of who pays – e.g. the EPSRC or the AHRC? where do you publish, where does it count in the research assessment exercise? These present questions present obstacles within the system which work against cross-domain sharing.

When asked more generally about data sharing and what funders should do, our interviewees highlighted:

11. Give advice and guidance
12. Make data available to people
13. Think then about data preservation planning.

The latter may best be done by handing over to an appropriate archive [or repository], or it may stay in the original institution, but in that case the institution will need:

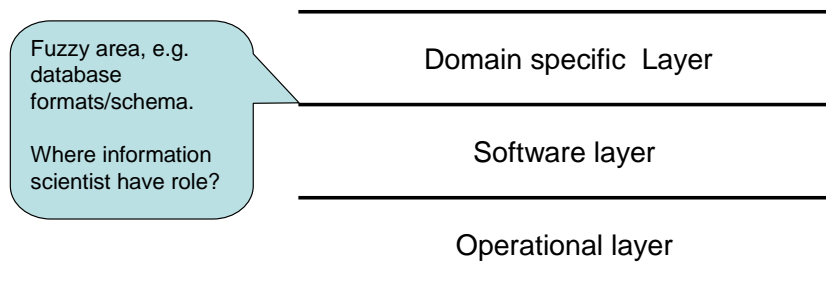
- i. Support, and
- ii. An exit strategy, for when and if they want to move data on to another storage facility.

The history community provides a good example of a community where data sharing used not to be the norm. Historians have little history of doing collaborative work, most efforts being by individuals – an historian's output is generally single-author books and papers. In some areas there are very small communities – an example given was that of researching electronic poll books. There were just four people in this area, all of whom know and talk to each other: formal sharing mechanisms were not relevant to them. The social science end of history is better at sharing than the others; an example is the use made of the 1800 census data, where this has been combined with other data from the USA, Canada etc, producing a combined, shared database on 98 million people. Published materials suffer from a "we are doing it right" attitude and possessive ownership attitudes. One is more likely to be offered research data – there has been more investment in creating it.

On the whole in the past historians have often been reluctant to use other historians' data, or the same materials. While the technical skills among them can be low (some are still using typewriters), the AHDS has seen a development in the skills sets, in part fostered by the AHDS' own services, a development which in turn is increasingly affects trends in usage of data.

When asked about data archiving and preservation in general, rather than just applied to the arts and humanities, an interviewee suggested that we are using old and inappropriate models (or examples), which are based on the libraries, archives, and publishing world. We need new terminology and models – or indeed perhaps a new theory to underpin the work being done.

Discussing where specific tasks should take place, it was suggested the following layered model be used. This has a software ingest and access layer on top of a basic operational and storage layer. On top of these there needs to be a domain-specific layer of support activities requiring knowledge of content and intent:



Between the domain specific layer and the software layer there was a fuzzy area concerned with such issues as database schema, and formats. In this area formal information scientists may have a role to play.

An interviewee also noted that we need to distinguish what is truly digital and what is not. What is a bona fide digital resource?

Another interviewee noted that activities in this area (preservation, curation,) need to cut across conventional categories - the research that needs to be done in this area are into selection and retention (perhaps an area for a body such as the Digital Curation Centre?). There is also a need to do research into risk assessment applied to data retention.

Some final points

The group noted that as a whole there is perhaps a tendency to neglect those who are outside the HE/FE (higher/further education, broadly “education”) arena – how are those in life-long learning, schools, genealogists served?

The group felt that there was a need for co-ordination in this area, a role which a body such as the Digital Curation Centre could fulfil.

Appendix 3.2 British Atmospheric Data Centre and NERC Data Grid

Key data

Name: British Atmospheric Data Centre
Acronym: BADC

Established: 1994

Brief description: The British Atmospheric Data Centre (BADC) is the Natural Environment Research Council's (NERC) designated data centre for the Atmospheric Sciences. The role of the BADC is to assist UK atmospheric researchers to locate, access and interpret atmospheric data and to ensure the long-term integrity of atmospheric data produced by NERC projects.

The NERC Data Grid (NDG), of which the BADC is to form a part, is to be a distributed system spread across the NERC data centres. The proposal is to build a grid which makes data discovery, delivery and use much easier than it is now, facilitating better use of the existing investment in the curation and maintenance of quality data archives.

Contact: Dr Bryan Lawrence, Director

British Atmospheric Data Centre
 Space Science and Technology Department
 R25 – Room 1.106
 CCLRC Rutherford Appleton Laboratory
 Fermi Avenue
 Chilton, Nr Didcot, Oxfordshire
 OX11 0QX.

BADC: <http://badc.nerc.ac.uk>

NDG: <http://ndg.nerc.ac.uk/>

Funding: Primary funding is by NERC.

Staffing: 12 in total

Data volume held The Centre manages many Terabytes of data.

Interview data: Bryan Lawrence (Director), on 3 November 2004 by Philip Lord and Denise Ecklund.

BADC Case study report

The British Atmospheric Data Centre (BADC) is one of seven data centres managed by NERC (see list in Addendum 1 below). It hosts some 110 data sets. The centre also hosts some 17 collaborative project workspaces, most of which are closed to public access.

BADC stores and manages data derived mainly from:

- NERC-funded research projects from the NERC thematic research programmes (when atmospheric data is involved); BADC has no knowledge of the data produced by NERC-funded “minor programmes” and “collaborative programmes”
- British Meteorological Office (the “Met Office”)
- European Centre for Mid-range Weather Forecasting (ECMWF), strictly under licence.
- ENVISAT data (from the European Space Agency’s newest satellite). ENVISAT raw data is duplicated at a small number of other European sites.

The initial data results derived by the original users/creators are not normally stored on the BADC system; typically such user-derived data remains with the individual users. When a NERC-funded atmospheric science project completes, the ‘final data results’ from that project must be submitted to BADC for archiving.

Data is transferred to BADC either on storage media or via network upload; they would prefer all incoming data to be transferred via the network. One full-time engineer is employed to load all incoming data into the BADC storage systems. There are no quality control checks on content when data is loaded, but they do check the metadata which is received – and users’ metadata is often poor-quality.

Based on use statistics gathered at the centre, BADC has more than 6,000 registered users worldwide, of which more than 1000 access the data centre in any one year. Users include members of the NERC atmospheric research community.

BADC data and services are accessed through the BADC website. The website maintains a catalogue of all datasets and tools. Datasets are divided into two categories:

- Unrestricted – these are openly available to the public via the website using FTP or a web browser interface.
- Restricted – these require the user to register with the BADC and to apply for access. Some of these are only available to bona fide academic researchers, and some require agreement to conditions of use.

The data is held as a collection of discrete files/databases in various formats – ASCII and binary. The on-line data catalogue provides full records of the data, and each resource also has a descriptive web page associated with it. The BADC also provides a third-party software catalogue for data users. Data mining facilities across and within datasets are not provided by the BADC – everything is done on a per data-set basis. Searching within data sources at the moment would require users taking the whole datasets needed themselves, or subsets of them if necessary, and then mining them.

Although currently data is transferred to users using standard FTP, in 2006 data transfer will be available using sophisticated FTP serving (using parallel transfers and auto-restart in the event of failure) to address the issues of transferring large data files. This will also require a high-speed internal network (2 Gbits per second) to move large data-sets.

The group run a help desk for users, and their web site provides extensive guidance and is easy to use.

The BADC has a research group attached to it and a science support team for the NERC thematic programs. This team produces requirements specifications for data management for NERC research projects.

Evolution of the resource and the NERC Data Grid

The BADC was established in 1994 primarily as a change in remit and status of the older Geophysical Data Facility.. It is funded solely by NERC. Data holdings have been acquired from other sources which go back decades, but the NERC datasets have only been acquired since the 1990's.

NERC created funding for the NERC Data Grid (NDG) project in 2003. The project will use Grid technologies to tie together many geo-data centres for cross- geo-disciplinary research. The BADC is included in the NERC Data Grid (NDG).

The NDG will support cross- geo-disciplinary research work by providing a single uniform access point to all the NERC data centres and to other geo-data centres around the world, but data sharing at this level will not be possible until the NDG technology is fully deployed.

BADC is in discussions with the National Center for Atmospheric Research (NCAR, USA) and the World Data Center on Climate (WDC, Max Planck Institute, Germany) to start a project to combine their data and tools catalogues, and to support reciprocal access. Again, NDG technology is expected to provide the technology and infrastructure necessary to share data and tools among these organisations.

Standards

Standards defined by the geo-science community focus on metadata models, controlled vocabularies, cross-mapping between vocabularies and other discipline-specific topics.

The geo-science community is not attempting to define standards for a general infrastructure, but the BADC is collaborating with computer scientists in projects such as the NDG which might impact general standards, based on their experience with the general data sharing requirements of the geo-science community.

The BADC uses geo-science standards whenever possible. They actively use the following:

- NASA Directory Interchange Format and protocols (DIF)
- Geography Mark-up Language (GML) - under development in ISO as ISO/CD 19136. NDG is working with an Australian group to extend GML to handle three-dimensional atmospheric data.
- Open Geospatial Consortium (OGC) standards for web map services, web coverage services and web features services
- Vocabulary mapping defined by the BODC, developed under the National Oceanographic Research Consortium. BADC does not define a similar vocabulary for atmospheric data, but it does promote and follow a standard naming convention for metadata values.

BADC uses several general-purpose standards from the web and grid communities, including ISO standards (including UML to define metadata models, XML schema for metadata models), W3C-compliant web services, and RDF (Resource Description Framework) to describe semantic relationships.

NDG is also a strong promoter of existing metadata standards, but is also extending these as necessary. NDG defines metadata standards and protocols for managing metadata. The primary goal is to enable 'cross-domain' research projects that use data formats they would not normally understand.

NDG is defining four types of metadata:

- A [Archive]: Format and usage metadata. Supports physical access to archived data
- B [Browse]: A superset of discovery, usage data, and contextual metadata. It supports browsing and selecting datasets
- C [Comment]: Annotations, documentation and other supporting material, including comments from the data creator.
- D [Discovery]: Metadata used to locate datasets. Supports data discovery (conformant with ISO 19115, which defines the schema required for describing geographic information and services)

There are no real standards for atmospheric data ontologies as yet. BADC commissioned a study to search for atmospheric data ontologies to support climate forecasting. They are awaiting the final report.

Neither are there standard vocabularies for atmospheric chemistry. BADC has 35,000 atmospheric chemistry data files; as yet they do not have a vocabulary to describe them in associated metadata they cannot make them available to researchers.

BADC are gradually framing their archiving work in terms of the Open Archival Information System standard (OAIS, ISO 14721), and indicated that they would support a DCC-led certification programme for standards.

BADC feels it does not necessarily require a body with the weight of ISO to establish standards. A group such as the Global Organization for Earth System Science Portal (GO-ESSP) may also establish international standards effectively.

GO-ESSP is a collaboration designed to develop a new generation of software infrastructure that will provide distributed access to observed and simulated data from the climate and weather communities. GO-ESSP will achieve this goal by developing individual software components and by building a federation of frameworks that can work together using agreed-upon standards. The GO-ESSP portal frameworks will provide efficient mechanisms for data discovery, access, and analysis of the data.

GO-ESSP promotes the establishment and adoption of standards. Three standards are currently in active use:

1. FGDC/ISO – for metadata schema. The group is working to ‘harmonize’ the Federal Geographic Data Committee (FGDC) metadata standard and the newly adopted ISO 19115 metadata standard. ANSI has recommended US adoption of ISO 19115. The FGDC group has defined a ‘cross-walk’ between the existing FGDC standard and ISO 19115.
2. DIF (Directory Interchange Format) – a standard for managing and exchanging metadata catalogue information among multiple organisations. The standard was defined in 1987 and has been extended to fully support the ISO 19115 standard adopted in June 2004. Only 8 fields are required in a DIF entry.
3. NetCDF Climate and Forecast Metadata Convention (CF) – This is a de facto standard. The standard defines semantics of metadata for climate and forecasting data stored in Network Common Data Form (NetCDF) file format. It also defines ‘standard names’ for units of measure. A CF-checker has been developed as supporting software. The CF community are considering whether their convention can be applied to data stored in the HDF format.

Tools

The BADC provides a suite of common data analysis tools and models, working with its research community to determine what applications and tools are needed. They offer web-based access to tools and data. (There is interest in comparing estimates/models between the ECMWF and Met Office.)

Tools for searching and selecting a data subset are very limited. Users can extract a subset of a file if they know and understand the file format. Otherwise, users must download entire datasets and trawl through the dataset to locate the relevant data. Selecting a data subset would be most helpful when the user wants to access only the changes and additions to the file since the last complete file download.

Whilst BADC developed its own browser-based tools to select BADC datasets and analysis applications, it is incrementally adopting NDG-developed tools as they become ready for deployment.

The geo-analysis applications were developed either by BADC or by other researchers who have agreed to make their tools available.

Open-source software is viewed as the best way to speed progress.

Budgeting, licensing, costs

The NERC board mandates that 5% of science budgets should be available for projects to address data management issues. However, the board cannot always force the NERC thematic committees to invest the 5% of their budgets in data management. All BADC data is supplied free of charge, whether the data is restricted or not. Could data centres like BADC charge for data access and services to become self-funded or to mitigate costs to NERC? BADC has investigated the ‘revenue potential’ for BADC services from non-NERC users. They concluded that:

- The cost of the commercial infrastructure would exceed anticipated revenues.
- After-sales support costs would be marginal.
- BADC stores historical data, commercial interest in atmospheric data is focused on weather prediction based on current and near-current atmospheric conditions. Therefore there is little market for the data.
- BADC only provides atmospheric data; commercial interest might be greater for geo-data crossing multiple sub-areas of the discipline.

In a discussion about the NERC model of long-term funding, for centralised data centres organised along subject lines a good model for storing and sharing life-science data, it was noted that whilst the data centre staff are trained and ready, it is difficult to support research projects that cross different data-centre boundaries. It is hoped that the NERC DataGrid can address these difficulties.

Legal and IP questions

- BADC acknowledge the need to tighten their current licence terms regarding legal liability. For example, oceanographic tide data may prove to be inaccurate and a boat owner following this information may run aground and damage the boat. While atmospheric examples are not obvious, BADC must take steps to ensure no liability can be attached to them in such cases.
- Security presents more of an issue for metadata than for the raw data or the data products created by BADC. For example, in the NDG context, revealing the location of sites from which crop samples have been taken may be inappropriate.

Intellectual property (IP) is probably not an issue for BADC’s atmospheric data per se. The IP rights to an analysis model belong to the researcher or the researcher’s institution. Researchers are reluctant to release their models because:

- They are then more open to analysis and criticism
- Their software might not be resilient and robust enough for use by others
- They do not believe that the sharing of data would be reciprocated.

Staffing issues

BADC employs 12 staff, who are engaged to develop tools, load data into the repository, and maintain the system in general. BADC also runs a help desk providing direct user support.

Finding the right staff is a problem. There is no formal data management career structure, and the work requires knowledge of the atmospheric sciences discipline as well as of IT. Staff retention is also an issue in that promotion committees may not be aware of the skills that are required for a given role. Part of this problem is that digital data management is a new occupation, although it has some analogies with those who maintain web pages:

- The data centre typically exists within a larger organisation, and promotion and review criteria are defined to suit the majority in the organisation
- Often these criteria are inappropriate for a data manager, data curator or tools developers
- Although the ‘technical career track’ may reach a high level, data centre staff may fail to be promoted due to inappropriate evaluation criteria.

The qualifications of the person are also important: ... he or she needs to be able to understand the discipline and be a good programmer. It is easier to teach the computer science than atmospheric science.

As commercial computer courses tend to be too expensive for the academic sector, the BADC does training itself. As an example, one way it did this was to organise a seminar series to import computer software expertise from other related groups at RAL (NDG, CCLRC, etc.). NIESS (National Institute for Environmental eScience in the Cambridge e-Science Centre) have run one or two training courses for BADC staff. BADC staff gave seminar courses to the NDG staff. A course on the Python programming language was organised and given internally.

Recent changes in employment regulations have some positive impact in this area. The new regulations promote retention of existing personnel if any appropriate jobs are available within the broader organisation. If an employee has been employed for more than three years (under a fixed-term contract), that employee is considered “permanent”. When the current contract lapses the employee can request and assume any other available job in the organisation. If no positions are vacant or the staff member is not qualified for any vacant position then the staff member can be made redundant. This gives staff the possibility to move to another position within the organisation.

Data sharing

To be able to share their data outside the atmospheric sciences community, users would be in for a lot of work – the data is in arcane formats which would need to be understood and processed. Within the discipline these formats are well understood. (But it was also noted that older data may have only a small community of knowledgeable users – such as data from old satellites.)

In particular it is difficult for life science researchers to use BADC data. Dr. Lawrence estimates that at least one year of data management work would be required to integrate some of the BADC data so that life science researchers could identify and extract data in these new and unfamiliar data formats.

Barriers exist at all levels, from social down to data – except at the institutional level where NERC has mandated deposit of data (though the deposit of data models and software used are not mandated – these will be owned by the individual or his/her institution).

The BADC talks to other repositories, but this is usually at a “discipline level” rather than a “repository level”. There are quarterly meetings of the seven data centres in NERC.

Metadata issues

The creation of metadata is a people problem because:

- It cannot be fully automated
- It is a recurring human labour-intensive activity that must be paid for each time new data is created.

Metadata is an annoyance to contributing users, and often for them it is an ownership issue, since metadata is considered by them to be theirs and gives them an advantage. As noted above, data creators (researchers) provide poor quality metadata, and they are not motivated to provide this information. They may not know reasonable values for much of it.

The BADC is trying to improve the quality of metadata:

- By ensuring early provision, where the researchers must assign metadata values early, as data is created;
- Once data and its associated metadata have been submitted to the archive, data access is restricted for some period of time, allowing exclusive use by the data creator for this time period.

Both these actions are within BADC’s sphere of influence. A similar approach is being taken by the World Data Centre on Climate. Curators at this centre screen and check the quality of the metadata, and they enforce that the metadata must always be provided.

The BADC would like to see additional incentives for researchers to improve the quality and quantity of their metadata. One approach is to include data citations in research publications and for employers and review committees to give research credits for data citations. This model is being supported by the American Geophysical Union (AGU), an international society of geophysical researchers. One could argue sharing enhances one’s academic standing. The possibility of dataset citation will help change the culture (and there was a German example of this). It is now possible to cite data in the AGU publications.

The BADC is also trying to obtain metadata for legacy datasets and new datasets. Creating metadata for legacy datasets is very difficult as the data creators are often unavailable. BADC prioritises the datasets and does its best to provide reasonable and accurate metadata.

When asked what is the biggest hole to be filled, or what priority funding areas were there, the answer was tools (including policies) to handle metadata, and incentives to create metadata.

Key messages to the study sponsors

We asked what the key messages would be to this study's sponsors. The reply was:

Preserving data is going to be more expensive than sponsors and funders probably think it will be. Data preservation is a relationship between data producers and data consumers. Tough decisions must be made about what should be preserved. This is made more difficult by the fact that current economic market models and current intellectual market models are inappropriate for making decisions about long-term data preservation.

All NERC data centres are likely to receive long-term funding as they provide primary data services to NERC-funded researchers and to the community at large. The data held by the centres should be available in perpetuity. This is difficult to achieve. Old data formats are difficult to deal with. Typically an old format is known by fewer people. If no one knows the format of an old dataset, then the easiest option is for that dataset to be discarded.

They recommend the creation of an 'Electronic Digital Curation Journal'. The journal would hold links to datasets, metadata and dataset authors. The journal should be compatible with the ISI Citation Index so it could be valued in the RAE.

Addendum 1 – NERC Data centres

NERC funds one data centre for each of seven thematic research areas. Three are large data centres: BADC, BODC and British Geological Survey. These centres provide data services to NERC-funded researchers and the geo-communities at large. The other centres are smaller and primarily provide services only to NERC-funded researchers in their respective topic areas.

Acronym	Data Centre (and link)	Domain covered
AEDC	Antarctic Environmental Data Centre	Polar Science
BADC	British Atmospheric Data Centre	Atmospheric Science
BODC	British Oceanographic Data Centre	Marine Science
EIC	Environmental Information Centre	Terrestrial & Freshwater Science
NDGC	National Geoscience Data Centre	Earth Sciences
NWA	National Water Archive	Hydrology
NEODC	NERC Earth Observation Data Centre	Earth Observation

The NERC's science-based archaeology community is encouraged to deposit data with the AHDS Archaeology Data Service (see the AHDS case study).

Appendix 3.3 Biomedical Research Informatics Delivered by Grid Enabled Services

Key data

- Name:** Bio-medical Research Informatics Delivered by Grid Enabled Services
- Acronym:** Bridges
- Brief description:** BRIDGES is two-year a collaborative project developing and exploring database integration over six geographically distributed research sites within the framework of the large Wellcome Trust's Cardiovascular Functional Genomics (CGF) biomedical research project. Three classes of integration are being developed to support a sophisticated bioinformatics infrastructure supporting: data sources (both public and project generated), bioinformatics analysis and visualisation tools, and research activities combining shared and private data. Both OGSA-DAI and IBM Information Integrator technology are being employed and a report will identify how each performed in this context. A key focus of the work has been to address security issues also. The scientists have their own research and experimental data sets (microarray expression data and QTL data sets) that they wish to share *only* with one another.
- Contact:** Professor David Gilbert, Professor of Bioinformatics;
Bioinformatics Research Centre
A416 Davidson Building
University of Glasgow
Glasgow G12 8QQ
Scotland, UK

<http://www.brc.dcs.gla.ac.uk/~drg>
- Funding:** Primary funding is by DTI.
- Report data:** Dr Richard Sinnott, National e-Science Centre, University of Glasgow
Interviewees by Richard Sinnott, Denise Ecklund also with Derek Houghton, Neil Hanlon of BRC

BRIDGES Case study report

1. Introduction

The BRIDGES project [A3.3 1] is a prime candidate for a case study used to explore the practical issues involved with data sharing in the life science domain. The BRIDGES project, which began in October 2003 and is due to run until December 2005, is a supporting project for a large Wellcome Trust biomedical research project: Cardiovascular Functional Genomics (CFG) [A3.3 2]. The CFG project is investigating the genetic causes of hypertension. The project is pursuing a translational strategy, combining studies on rodent models of disease with parallel studies of human DNA collections. The primary concerns of BRIDGES are the development of a Grid infrastructure which facilitates improved CFG life science research through integrating and/or federating a broad range and class of biomedical data sources, and providing a suite of bioinformatics tools exploiting this data, and as such tailored to investigating functional genomics research. This case study thus represents a concrete exploration into what can be achieved now in terms of access to and usage of a variety of life science data resources.

This case study is structured as follows. We begin with a general classification of the data that arises within the CFG project and briefly outline the Grid technologies that BRIDGES is applying to access and use the different life science sets. Section 3 then looks in detail at the specific data resources of interest to the CFG, and hence BRIDGES projects and provides an account of the experiences gained in accessing - or attempting to make use of these resources. Finally section 4 outlines discussions that have taken place with the users, data providers and curators, technology providers, standards developers associated with these data resources, and draws conclusions on possible ways forward to improve the overall data sharing.

2. Data Classification

The data of concern to the CFG project and hence to BRIDGES can be classified in various ways. We outline them here (as described in the original proposal) since they underpin several key issues in data sharing, namely understanding the origin, history, usage (implied and explicit) issues associated with potential data sharing. This classification is generic and readily applies to different data sharing application domains and contexts. For brevity “...” is used to omit unnecessary text.

- Public data: data from public sources, such as Swiss-Prot (now UniProt) and EMBL. These may be held as local copies ... and can be shared throughout the consortium and with other consortia.
- Processed public data: public data that has additional annotation or indexing to support the analyses needed by CFG. These must be held within the consortium, but one copy can serve the entire consortium. They may be of interest to and made available to other consortia.
- Sensitive data: ... require careful enforcement of privacy and may be restricted to one site, or even part of a site.
- Special experimental data: this may fall into a particular category, e.g. microarray data, which has special arrangements for its storage and access already agreed.
- Personal research data: data specific to a researcher, as a result of experiments or analyses that that researcher is performing. This is not shared even among the local team. It may later become team research data.
- Team research data: data that is shared by the team members at a site or within a group at a site. It may later become consortium research data, e.g. when the researchers are confident of its value or have written about its creation and implications.
- Consortium research data: Data produced by one site or a combination of sites that is now available for the whole consortium.
- Personalisation data: Metadata collected and used by the bioinformatics tools pertinent to individual users. This data is normally only needed to support the specific user to which it pertains. But it may need to move sites when bioinformaticians visit sites or work together.

As seen from this brief characterisation, data can be classified in many ways and it is important to realise the context from which the data originates and the “intended” usage of the data before one contemplates the technical issues in describing the data (or meta-data) in a form where it might be subsequently published, accessed/used, or the relevant standards that data might be expected to adhere to.

The above characterisation of data can also be broadly reformulated in terms of the security associated with the CFG data, i.e. it is either public, shared (at some level, e.g. between all partners, between some) or private. Security in the context of the Grid is often used to establish Virtual Organisations as (VOs) depicted in Figure 1. VOs provide a framework through which the rules associated with the participants and resources are agreed and enforced – especially those specific to security requirements, e.g. on data sharing. The distribution of CFG partners and the data security needs are depicted in Figure 1.

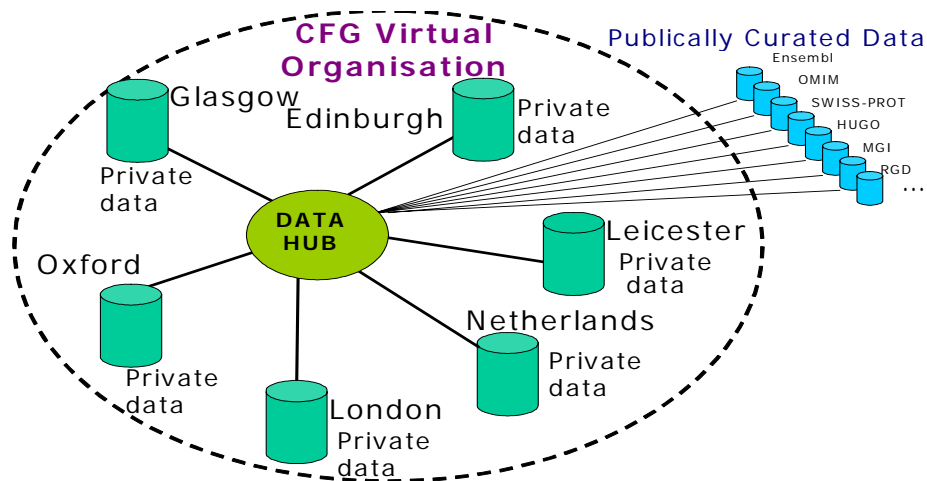


Figure 1: Data Distribution and Security of CFG Partners

A central component to this VO within the BRIDGES project is the notion of a Data Hub which provides the central data resource for the CFG scientists. The Data Hub itself is a DB2 [A3.3 42] repository which makes use of two key technologies: the Open Grid Service Architecture Data Access and Integration (OGSA-DAI) middleware and a commercial data access solution, IBM's Information Integrator. OGSA-DAI and Information Integrator have broadly similar aims: to connect distributed heterogeneous data resources. The focal point of this case study is the extent to which the Data Hub has allowed scientists *effective* access to the relevant data sources. The term "effective" has been italicised here since many of the data resources are public and can be accessed and used by the scientists right now, e.g. through a web page which allows queries to be issued against the data itself. Whilst effective for an individual resource, this approach is not conducive to the wider problem of data sharing. Life science data sets do not exist in isolation. The ability to link a multitude of different data sets is a key part of the day to day activities of a life scientist – from sequence data right through to data sets on population demographics. With this simple web based model of data sharing, life scientists typically are involved in manual trawling of the internet, following a multitude of hyperlinks to various web (data) resources in the search for information (more data). As well as being a human-intensive activity, this model of data access is fraught with problems and often leads to fruitless navigations of the internet.

A better model of data sharing is needed that ideally provides access to the data resources themselves, where queries, e.g. using Structured Query Language (SQL), can be made which can directly use the various remote data repositories. This is one of the key areas of investigation for the BRIDGES project.

3. BRIDGES data sources

3.1 Public data sources

Based upon a requirements capturing exercise with the CFG scientists, it was identified that the following data resources were of most interest/relevance to the CFG scientists:

- ENSEMBL (rat, mouse and human genomes) [A3.3 46, 47]
- Online Mendelian Inheritance in Man (OMIM) [A3.3 48]
- Human Genome Organisation (HUGO) [A3.3 49]
- Mouse Genome Institute (MGI) [A3.3 50]
- Rat Genome Data Base (RGD) [A3.3 51]
- SWISS-PROT [A3.3 52]
- Gene Ontology (GO)

This is not a complete list since the BRIDGES project was still on-going. New data resources can and will be included as the project progresses. These data resources do however represent some of the most widely used and accepted public genomics data resources hence it is essential that these data sets can be shared “effectively” by life science researchers as described above.

3.2 CFG Data Sources

The following data resources have been agreed to be shared between the CFG consortia

- Microarray data sets
- Quantitative Trait Loci (QTL) data sets
- Affymetrix chip mapping data sets

4. References

Relevant Projects

- [A3.3 1] Cardiovascular Functional Genomics project,
<http://www.brc.dcs.gla.ac.uk/projects/cfg/>
- [A3.3 2] BioMedical Research Informatics Delivered by Grid Enabled Services (BRIDGES), www.brc.dcs.gla.ac.uk/projects/bridges

Security-related

- [A3.3 5] E-Science Security Roadmap: Technical Recommendations v0.5, UK e-Science Security Task Force, draft executive summary v0.51
- [A3.3 6] ITU-T Rec. X.509 (2000) | ISO/IEC 9594-8 The Directory: Authentication Framework
- [A3.3 7] C Adams and S Lloyd (1999), Understanding Public-Key Infrastructure: Concepts, Standards, and Deployment Considerations, Macmillan Technical Publishing.
- [A3.3 8] Adams, C., Lloyd, S. (1999). “Understanding Public-Key Infrastructure: Concepts, Standards, and Deployment Considerations”, Macmillan Technical Publishing, 1999
- [A3.3 9] Austin, T. “PKI, A Wiley Tech Brief”, John Wiley and Son, ISBN: 0-471-35380-9, 2000
- [A3.3 10] Grid Security, <https://forge.gridforum.org/projects/sec>

- [A3.3 11] L Pearlman, et al., A Community Authorisation Service for Group Collaboration, in Proceedings of the IEEE 3rd International Workshop on Policies for Distributed Systems and Networks. 2002.
- [A3.3 12] M Thompson, et al., Certificate-Based Access Control for Widely Distributed Resources, in Proc 8th Usenix Security Symposium. 1999: Washington, D.C.
- [A3.3 13] VOMS Architecture, European Datagrid Authorization Working group, 5 September 2002.
- [A3.3 14] Steven Newhouse, Virtual Organisation Management, The London E-Science centre, <http://www.lesc.ic.ac.uk/projects/oscar-g.html>
- [A3.3 15] D. Chadwick and A. Otenko. The PERMIS X.509 role based privilege management infrastructure, in Proceedings of the Seventh ACM Symposium on Access Control Models and Technologies, Monterey, California, USA. 2002.
- [A3.3 16] Privilege and Role Management Infrastructure Standards Validation project www.permis.org
- [A3.3 17] P Hallem-Baker and E Maler, Assertions and Protocol for the OASIS Security Assertion Markup Language (SAML), OASIS, SAML 1.0 Specification. 31 May 2002. <http://www.oasis-open.org/committees/security/#documents>
- [A3.3 18] I. Denley and S.W. Smith, Privacy in clinical information systems in secondary care. British Medical Journal, 1999. 318: p. 1328-1331.
- [A3.3 19] Platform for Privacy Preferences (P3P) Project, W3C, <http://www.w3.org/P3P/>

Relating to Grid Data Standards

- [A3.3 20] Data Access and Integration Services working group, <https://forge.gridforum.org/projects/dais-wg>
- [A3.3 21] Global Grid Forum, www.ggf.org
- [A3.3 22] Grid Data Service Specification, https://forge.gridforum.org/docman2/ViewCategory.php?group_id=49&category_id=517
- [A3.3 43] Global Grid Forum, Open Grid Service Infrastructure, <https://forge.gridforum.org/projects/ogsi-wg>
- [A3.3 44] Global Alliance, Web Service Resource Framework, <http://www.globus.org/wsrf/>
- [A3.3 45] Organization for the Advancement of Structured Information Standards, <http://www.oasis-open.org/home/index.php>

Technology-related

- [A3.3 23] Apache web site, www.apache.org
- [A3.3 24] Web Security Standards, http://www.oasis-open.org/committees/documents.php?wg_abbrev=wss
- [A3.3 25] UK e-Science Engineering Task Force, www.grid-support.ac.uk/etf
- [A3.3 31] Basic Local Alignment Search Tool (BLAST), <http://www.ncbi.nlm.nih.gov/Tools/>
- [A3.3 3] Open Grid Service Architecture – Data Access and Integration project (OGSA-DAI), www.ogsadai.org.uk

- [A3.3 4] IBM Information Integrator,
<http://www3.ibm.com/solutions/lifesciences/solutions/InformationIntegrator.html>
- [A3.3] e-Science Data Information and Knowledge Transfer (eDIKT) – www.edikt.org
- [A3.3 41] Replica Location Service (RLS), www.globus.org/rls
- [A3.3 42] IBM DB2 Universal Database, <http://www-306.ibm.com/software/data/db2/udb/>

Relating to Life Science Data Resources

- [A3.3 29] EMBL-EBI European Bioinformatics Institute, <http://www.ebi.ac.uk/ensembl/>
- [A3.3 35] NCBI GenBank, <http://www.ncbi.nlm.nih.gov/Genbank/>
- [A3.3 36] NCBI LocusLink, <http://www.ncbi.nlm.nih.gov/LocusLink/>
- [A3.3 37] NCBI Online Mendelian Inheritance in Man,
<http://www.ncbi.nlm.nih.gov/OMIM/>
- [A3.3 38] NCBI Unigene, <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene>
- [A3.3 39] PubMed Central Home, <http://www.pubmedcentral.nih.gov/>
- [A3.3 40] US National Library of Medicine, <http://www.nlm.nih.gov/>
- [A3.3 42] EMBL-EBI European Bioinformatics Institute clustalw,
<http://www.ebi.ac.uk/clustalw/>
- [A3.3 43] EMBL-EBI European Bioinformatics Institute MPSrch,
<http://www.ebi.ac.uk/MPsrch/>
- [A3.3 46] ENSEMBL Trace Server, <http://trace.ensembl.org/>
- [A3.3 47] ENSEMBL Trace Browser, <http://www.ensembl.org/>
- [A3.3 48] Online Mendelian Inheritance in Man (OMIM),
<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>
- [A3.3 49] Human Genome Organization (HUGO), <http://www.gene.ucl.ac.uk/hugo/>
- [A3.3 50] MedLine, <http://www.ncbi.nlm.nih.gov/PubMed/>
- [A3.3 51] Mouse Genome Informatics (MGI), <http://www.informatics.jax.org/>
- [A3.3 52] Rat Genome Database (RGD), <http://rgd.mcw.edu/>
- [A3.3 53] Swiss-Prot, protein knowledgebase, http://us.expasy.org/sprot/sprot_details.html
- [A3.3 54] Affymetrix database,
<http://affymetrix.arabidopsis.info/narrays/images/NASCArraysTutorial.pdf>

Addendum – Data Sharing Technologies

OGSA-DAI

The Grid community is currently developing appropriate specifications for data access and integration in a Grid environment through the Data Access and Integration Service working group [A3.3 20] at fora such as the Global Grid Forum [A3.3 21]. Much of this work is driven by results from the OGSA-DAI project [A3.3 3] and the recently funded follow up project, Data Access and Integration 2 (DAIT) [A3.3 3]. OGSA_DAI/DAIT is a collaborative programme of work involving the Universities of Edinburgh, Manchester and Newcastle, the National e-Science Centre, with industrial participation by IBM and Oracle.

Their principal objective is to produce open source database access and integration middleware which meets the needs of the UK e-Science community for developing Grid and Grid related applications. Its scope includes the definition and development of generic Grid data services providing access to and integration of data held in relational database management systems, as well as semi-structured data held in XML repositories.

OGSA-DAI have focused upon making these data resources available within an OGSA compliant architecture. The OGSA-DAI Grid services themselves provide the basic operations that can be used to perform sophisticated operations such as data federation and distributed queries within a Grid environment, hiding concerns such as database driver technology, data formatting techniques and delivery mechanisms from clients. This is achieved by the provision of a Grid-enabled middleware reference implementation of the components required to access and control data sources and resources.

OGSA-DAI itself can be considered as a number of co-operating Grid services. These Grid services provide a middleware layer for accessing the potentially remote systems that actually hold the data, i.e. the relational databases, XML databases or, as planned for the near future, flat file structures. Clients requiring data held within such databases access the data via the OGSA-DAI Grid services. The precise functionality realised by OGSA-DAI is described in detail in the Grid Data Service Specification [A3.3 22]. A typical scenario describing how this functionality might be applied to find, access and use (remote) data sets involves a persistent DAI Service Group Registry (made available via a Grid services hosting container such as Apache Tomcat [A3.3 23]) offering persistent service factories (used for creating services to access and use specific data resources). Clients would contact the DAI Service Group Registry to find out what data sets are available, and once a required data source was found, create an instance of the Grid data service (via the appropriate factory) that would give access to this resource. The client can then issue queries (submit Perform operations via XML documents) to this Grid data service which extracts the queries and submits them to the appropriate databases, e.g. as SQL queries, before results are finally returned in XML documents. Extensions to this scenario to have multiple Grid data services supporting multiple, parallel queries executing through a given client query are possible.

It should be noted that the specifications and hence implementations such as OGSA-DAI are evolving continually. In March 2004 it was announced by the GGF that the previous approach for modelling Grid services in an OGSA framework – termed Open Grid Service Infrastructure (OGSI) [A3.3 43] would change to be more purely web service based. This new specification has been termed Web Service Resource Framework (WSRF) [A3.3 44] and is currently being elaborated on at fora such as GGF and Organization for the Advancement of Structured Information Standards (OASIS) [A3.3 45].

Information Integrator

Information Integrator – which was previously known as DiscoveryLink - has been developed to meet the challenge of integrating and analyzing large quantities of diverse scientific data from a variety of life sciences domains. IBM Information Integrator offers single-query access to existing databases, applications and search engines. The Information Integrator solution includes the combined resources of Information Integrator middleware and IBM Life Sciences services. Using this software, IBM Life Sciences services can create new components that allow specialized databases—for proteomics, genomics, combinatorial chemistry, or high-throughput screening—to be accessed and integrated quickly and easily.

Information Integrator talks to the data sources using wrappers, which use the data source's own client-server mechanism to interact with the sources in their native dialect. Information Integrator has a local catalogue in which it stores information (metadata) about the data accessible (both local data, if any, and data at the backend data sources). Applications of Information Integrator manipulate data using any supported SQL API, for example, ODBC or JDBC are supported, as well as embedded SQL. Thus an Information Integrator application looks like any normal database application.

The focus of OGSA-DAI and Information Integrator is primarily upon access and integration of data and not specifically upon security concerns. Security in the context of the Grid is an area that is currently receiving much attention since it is a crucial factor in the wider uptake of the Grid and with regards to this study, for data sharing and all of its ramifications. There are numerous standards under development addressing aspects of security [A3.3 10,24]. As such, both OGSA-DAI and Information Integrator largely focus on data access and usage where direct access to a database is given, i.e. via a programmatic API. This is predominantly not the case in current cases, even with public data resources such as those that exist in the life science domain.

Appendix 3.4 Clinical e-Science Framework

Key data

Name:	Clinical e-Science Framework
Acronym:	CLEF
Established:	October 2002 –December 2007 (CLEF project and CLEF-services combined)
Brief description:	CLEF is an MRC-sponsored project in the e-Science programme that aims to establish policies and infrastructure for the next generation of integrated clinical and bioscience research. A wider ambition is to bring the system into wider use by following the project phase with a CLEF-services phase. A major goal of the project is to provide a pseudonymised repository of histories of cancer patients that can be accessed by researchers. Robust mechanisms and policies are needed to ensure that patient privacy and confidentiality are preserved while delivering a repository of such medically rich information for the purposes of scientific research. Once evaluated, it is hoped that the CLEF approach can serve as a model for other distributed electronic health record repositories to be accessed for research.
Contact:	Professor Alan Rector, CLEF Project Director Kilburn Building Department Of Computer Science University of Manchester Oxford Road Manchester M13 9PL. http://www.clinical-escience.org http://clef-user.com/index.html
Funding:	Primary funding is by MRC.
Interview data:	Professor Alan Rector at the University of Manchester (Richard Sinnott, Alison Macdonald); Dr Dipak Kalra, Prof David Ingram at UCL (Philip Lord, Alison Macdonald); Dr Anne Westcott at AstraZeneca, (by telephone, Alison Macdonald, Philip Lord)

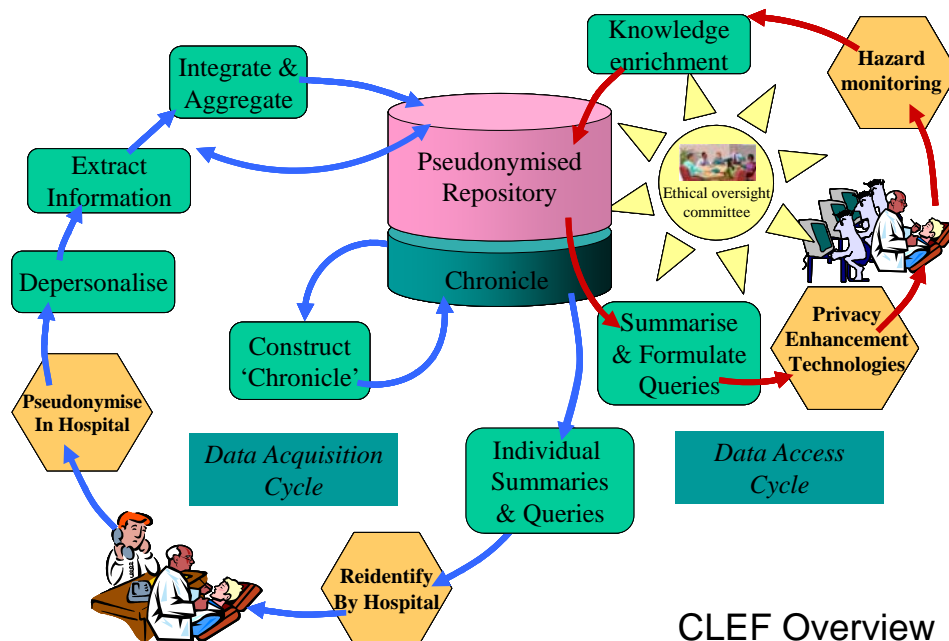
CLEF Case study report

CLEF overview

CLEF (Clinical e-Science Framework) is a three-year project sponsored by the MRC within the e-Science programme. This multi-million pound project began in October 2002, involving many partners and research participants (see below). The Project Director is Professor Alan Rector at the University of Manchester. Prof Alan Rector is a leading researcher in the area of clinical and life science data sets and the issues associated with their access and usage, (including Grid technologies). As well as being the PI for the Clinical e-Science Framework (CLEF) project, other relevant projects on which Professor Rector is working include the Prescription Indicators, CO-ODE, HyOntUse and the GALEN projects.

As CLEF is a very large project, highlights rather than a detailed description are provided here. It is concerned with enabling the sharing of clinical information not only between clinicians, but also between researchers, and others in the healthcare services. In doing this it makes the confidential nature of these records central to the design of the system. An excellent animated presentation is available on the web which describes clearly CLEF and its context (see: <http://www.clinical-esceince.org/main.html>); a short video is also provided.

The goal of CLEF is to provide a coherent, integrated framework of clinical information to, primarily, clinicians and to the research community using e-Science techniques. Central to CLEF is the establishment of a pseudonymised repository of electronic health records (EHR) (see figure). It draws on the *openEHR* medical records work by Professor David Ingram at University College London and colleagues. (See: <http://www.openehr.org/>)



CLEF Overview

This repository sits at the intersection of two data flows; that for the creation, pseudonymisation and use of clinical records (data acquisition cycle), and the research cycle (data access cycle). The latter only provides access to anonymised data (and includes facilities for identifying and monitoring risks to confidentiality) and is overseen by an ethics committee. (Re)-identification of patients in the clinical data flow is possible, with safeguards on privacy. Protecting confidentiality and security is a major part of the study (40% of CLEF budget goes to the security area).

Text extraction is a main area of work. Text records are at the heart of clinical information; in CLEF text analysis and processing for extraction and codification of information as well as anonymisation is a central function of the system. CLEF is also working to “disambiguate” information, that is to differentiate and clarify ambiguities or differences in the meaning of words.

From the pseudonymised repository a “chronicle” is created for each patient’s history and made available for access. The *chronicle* is intended to provide an element of understanding of the clinical process not only of *what* happened, but also *why* it happened. The system also allows the import of data from multiple sources; these are not necessarily textual - such as clinical images, laboratory results and tracings from measuring devices, etc.

The project has focused initially on using cancer records, as cancer is a priority for research and the NHS, it is seen as important by the public, and gene research is strongly linked to cancer. Some 22,000 records relating to deceased patients only have been loaded into the system from the Royal Marsden Hospital. The system is extensible to other clinical domains.

Project participants

The CLEF project is being undertaken by seven partners:

- University of Manchester, Medical Informatics Group (Prof Alan Rector)
- University of Brighton, Information Technology Research Institute (Prof Donia Scott and Dr Richard Power)
- University of Cambridge, Judge Institute for Management Studies (Prof Don Detmer and Dr Peter Singleton)
- University College London, CHIME (Prof David Ingram and Dr Dipak Kalra)
- Cathie Marsh Centre for Census and Survey Research (Dr Mark Elliot)
- University of Sheffield, Natural Language Processing Group (Dr Rob Gaizauskas and Dr Mark Hepple)
- Royal Marsden Hospital, Information Directorate (Dr. Jo Milan)

Goals and the longer term

The project has summarised its goals as follows:

- Enable secure and ethical collection of clinical information from multiple sites
- Analyse, structure, integrate, prioritise and disseminate

- Make resources available using GRID tools (e.g. myGrid)
- To provide access to clinical data to authorised clinicians and scientists
- Establish a secure and ethical framework for scientific collaboration

They further define the desired outcomes/deliverables of the project as:

- Shared repositories integrating clinical, image and genomic information linked to medical records, literature and the Web.
- Information capture and language tools for clinical and scientific descriptions to populate the repositories and records.
- Information integration tools to bring together information from isolated studies and disparate medical records in the repositories and to facilitate constructing future studies and records to integrate smoothly into the repositories.
- Information display tools to make the information easily accessible, including where appropriate to patients and public.
- Knowledge management environments (the Clinical e-Science Workbench) to use the repositories to enhance biomedical research, based on generic architectures, and contributing to the relevant clinical and technical standards.

A wider ambition is to bring the system into wider use, and to this end a group of NHS, commercial (IT, publishing and pharmaceutical) and academic collaborators and supporters are being established. Commercial collaborators include AstraZeneca, GlaxoSmithKline, IBM, Oracle and Sun.

In the longer term the project and its sequel aims for CLEF to become a key player in global healthcare informatics, and to maximise benefits from R&D projects via partnerships.

Conclusions from interviews

This project has first-hand experience into the issues into accessing and using such data sets. A key problem faced by the project has been in gaining ethical permission into accessing these resources.

The CLEF team and the Multi-Centre Research Ethics Committee (MREC) which evaluated CLEF reached a compromise agreement that only records of deceased patient would be used initially. However, indications are that only an ‘express consent’ approach would be deemed acceptable for access to living patients’ records, which CLEF considers infeasible, excessive to any risk involved, and not required under the Data Protection Act 1998. This has then led them into extensive activity to promote wider debate and to lobby for consideration of the issues at a policy level. ..

Despite this, the project has been unable to access and use live clinical data sets. The focus has been up until now to consider data sets from deceased patients, minimising data protection problems. The usefulness of this information in clinical trials, e.g. exploring the suitability of certain drugs over a given cohort is, of course, thereby substantially reduced. It limits the usefulness and pertinence of the data, restricting consideration to older cases (when

there were lower success rates or unsuccessful treatments). This does allow for explorations into needed features, e.g. services for anonymisation, however. Further it was noted that the development of clinical trials does not typically allow for re-use of services/infrastructures. Instead the data is collected under carefully controlled conditions to answer specific questions given by the nature of the trial. One of the main benefits of CLEF has been to establish collaborations (channels of communication) with the clinical data set providers/curators, as opposed to explicit technological solutions that solve problems in accessing and using these data sets. Procedural aspects of data creation and usage are a key facet to be understood before data itself can be shared.

The change in paradigm (with the advent of the genome completion) with more of an emphasis on *in-silico* research has also introduced further complexities for ethical bodies and data providers/curators. Which projects/people should their data sets be made available to, larger project such as UK BioBank or other projects such as CLEF? There is the possibility of allowing access to one “opening the door for others”. It was noted that these decisions have to be taken at national level initiatives and through agreed organisations, e.g. research councils, the government. Once this has happened, the technological solutions for data access, usage and security can be addressed. This is likely to be a slow and methodical process since there is much at stake – perhaps most significantly the level of trust between the various groups.

The project has spent a huge amount of effort in lobbying and discussing aspects over access to various clinical data sets. This was made much more difficult following the events at the Alder Hey Children’s Hospital in Liverpool, since when the whole issue of ethics in the medical domain has a significantly raised profile. One vision of how such issues might be overcome is a national consent register where patients may explicitly express in what context/under what circumstances and for what purpose their data may be made available, and importantly what exceptions to this might exist. This would help to overcome the many ethical committee barriers currently in place to safeguard the data sets.

An issue with the Grid in this domain is that it has to be “seen” to be a success by clinicians/researchers. Thus, is it the case that technology that allows access to data is likely to improve the research that is undertaken - traditionally measured by publications to *Lancet*, *Nature* and relevant journals etc? The example of myGrid¹⁵ was given as a success story since it is used to publish work in recognised journals (on Graves disease, Williams Beuren Syndrome etc). One worry about the medical/clinical domain was that it is often not possible to produce proof of concept systems that can be used to demonstrate the usefulness and applicability of Grid technology. For example, numerous demonstrations exist showing how public genomic data resources can be accessed and used through single optimised queries (via Grid technology). Such examples in clinical/medical domain are not so straightforward to achieve due to the limitations on data access. Another potentially more worrying issue in this domain is that open source solutions such as Grid will never be acceptable. There is always the possibility in non-validated and rigorously engineered systems for flaws to exist.

It was noted that medical data was generally high in value, since the whole process of medical data production was more human-intensive with emphasis on quality assurance, whereas biological data does not always have the same level of intensive input in its production.

¹⁵ Interestingly the technological solutions adopted by the myGrid project are not explicitly recognised as being Grid based – more web services.

Security was underlined as essential. The need to address and to be seen to address AAA (authentication, authorisation and accounting/auditing) is paramount to engage the medical community and the wider public. Confidence in security and acceptance of the wider framework, of which security is a part, are critical. If confidentiality has been promised, it is absolutely essential for that undertaking to be respected. Security is part of the consent process, in two ways: people must be able to have confidence in security, but they also accept risk in exchange for benefit – if the risks can be lowered, then it is easier for people to choose to be involved.

Even with this in place, there might well be a slow process of gaining trust from the NHS community and the outside world. Systems based upon “thin clients” (a computer-network device that can process information independently but relies on computer servers for applications, data storage, and administration) which allow access through firewalls are likely to be a phenomenon that will be around for some time between the NHS and Grid/academic communities.

Risk is always present in any information system (something which ethicists and the legal profession do not always recognise). It is important to be able to quantify risk and the likelihood of data being accessed by unauthorised users. Benchmarking security and/or assigning a security certificate based upon the procedures and technical solutions used to safeguard data and the infrastructure upon which it exists is one way that such issues could be addressed constructively. A national security body capable of reviewing security arrangements and practices would be of benefit – even outside the Grid/medical domains. For the commercial sector, in particular the pharmaceuticals industry, security is key to sharing and collaboration with the academic sector.

It was noted that there are several major initiatives/standardisation developments under way that are shaping clinical and medical data: Health Level 7 (HL7), SNOMED-CT, CEN and *openEHR*. HL7 (currently in version 3) is an international standardisation group. Whilst initially driven by the USA, the HL7 standards have significant acceptance by the NHS in the UK. One issue relating to HL7 was both the size and scope of the standards as well as the implementations of it. HL7v2 was implemented widely but had the problem of offering numerous “options”. This is an issue for data sharing since different systems implementing this standard may well not inter-operate (since they implement different sets of options). HL7v3 was put forward to overcome some of these issues. HL7v3 relies upon a Reference Information Model through which systems should map their data sets combined with a controlled vocabulary. Professor Rector had some concerns as to whether HL7v3 would be able to address all the interoperability issues, however. He also noted that everything on the NHS network has to be SNOMED-CT coded.

A key aspect of medical (and biological) system inter-operability is the use of terminology. Numerous groups are looking into this area, e.g:

- Gene Ontology (GO) which is widely regarded as a model for the way in which terminologies/ontologies can/should be put together
- MGED (Microarray Gene Expression Data)
- OMIM (Online Mendelian Inheritance in Man™)

Groups are already looking into this area, e.g. mouse embryo anatomy work at Edinburgh University (Mouse Atlas project) and mouse adult anatomy being done at MGI in Jackson. The University of Washington is also developing a digitised human anatomical resource.

The Unified Medical Language System (UMLS) began in 1990 with the intention to provide cross mappings between different terms in different species. This built on numerous linguistic sources with the idea of identifiers, e.g. Concept Unique Identifiers, Linguistic Unique Identifiers, String Unique Identifiers. The relation to Life Science Identifiers was noted. All of these efforts are trying to bring clarity/understanding and agreement to terms/phrases. This barrier is difficult in a single domain but exacerbated across domains. The example was made from the GALEN project of the term *neoplasia*. This term generally means “new growth for benign/malignant tumours in northern Europe” but “cancer in southern Europe”. Consensus reaching activities on names and their meaning *must* happen for data sharing to be effective even within a single domain. This has an impact on the education of the different communities. Combining information from different domains, e.g. medicine and biology is only possible if the different research communities have reached agreements in isolation.

Terminologies based upon taking the “union of everything are generally not useable by anyone” – they become too large, unwieldy. One successful approach applied in the GALEN project (Generalised Architecture for Languages, Encyclopaedias and Nomenclatures in medicine, see: <http://www.cs.man.ac.uk/mig/projects/old/galen/>) was to start from the bottom up. This project addressed multi-language challenges from the outset. Key to its success was the identification that the “concepts” of terms/phrases was of far more importance of words themselves. Conflict resolution was often a prolonged and time-consuming process.

A key issue was also the standardisation process itself. HL7 does not work to explicit deadlines in the standardisation- process (similar to Global Grid Forum), whereas SNOMED-CT does. Emphasis on what is implemented and widely used is often more compelling than what standards bodies produce however. The importance of HCI and general usability aspects in this process cannot be overstated.

Appendix 3.5 Common Data Access

Key data

Name:	Common Data Access Limited
Acronym:	CDA
Established:	1995
Brief description:	<p>CDA is a not-for-profit subsidiary of UKOOA (United Kingdom Offshore Operators Association) set up in 1995 to provide data management services to its members and to the UK oil industry. CDA is directed by a Council of members and is managed by a full-time management team. CDA facilitates collaboration between oil companies, service companies and regulatory authorities.</p> <p>UKOOA is the representative organisation for the UK offshore oil and gas industry. Its members are companies licensed by the Government to explore for and produce oil and gas in UK waters.</p>
Contact:	<p>Mr Malcolm Fleming</p> <p>CDA Limited 2nd Floor 232-242 Vauxhall Bridge Road London SW1V 1AU</p> <p>CDA also has an office in Aberdeen; the CDA DataStore operation is currently hosted by Landmark EAME Ltd., also in Aberdeen.</p> <p>CDA: http://www.cdal.com UKOOA: http://www.ukooa.co.uk/</p>
Funding:	Not revealed. The CDA is funded by the subscriptions from CDA's members in the oil industry.
Staffing:	Two people at the head office in London, two at the Aberdeen office and 10 full-time equivalent the data processing facility outsourced to Landmark EAME.
Interview data:	Malcolm Fleming (Chief Executive) and Martin Wadsworth of CDA and John McInnes (Deal Project Manager) of BGS, were interviewed by Alison Macdonald and Philip Lord, 1 November 2004.

CDA Case study report

Common Data Access Limited (CDA) is a not-for-profit subsidiary of UK Offshore Operators Association (UKOOA). It was set up as CDA Limited in 1995 after an industry interest group was formed to provide data management services to members and to the UK oil industry. The CDA facilitates collaboration between oil companies, service companies and regulatory authorities. UKOOA is the representative organisation for the UK offshore oil and gas industry. Its members are companies licensed by the government to explore for and produce oil and gas in UK waters. The organisation has links to NERC through the British Geological Survey (BGS). CDA is directed by a Council of members and is run by a professional management team.

The CDA has some 30 full members. The DTI and two data release agents (companies that manage the release of data once exclusive rights to the data have expired) are present as affiliate members. Members as of October 2004 are listed in the addendum below. The operation is funded by full members whose subscriptions give them access to data through two mechanisms described below, DEAL and DataStore, and also to other services. Subscriptions are based on the data and services to be accessed, based on the following principles:

For the CDA DataStore & Data Management Services:

- Simple model, based on operated well count
- Paid directly by CDA member

For the DEAL Data Registry

- From UKOOA members (based on their proportion of total oil and gas production)
- Also a DTI contribution

For both

- One payment, annually in advance

The benefit to a company of joining CDA is that they can dispose of all other versions of any operated data held on the DataStore and have immediate access to data from more than 8,000 wells/wellbores. There are commercial equivalents to DEAL; but all of DEAL is “free” once the funding subscription is paid.

The precise subscription levels are confidential, but the company operates on a not-for-profit basis and allows a rolled-over surplus, including a development fund of 5% - 20%. They now have almost 40 years of data.

Some major steps in the evolution of the resource were:

- Nov 1993 Industry interest group started
- Mar 1995 CDA Limited incorporated, and QC Data Ltd given a five-year contract to manage the data.
- Jun 2000 CDA shares purchased by UKOOA; DEAL service contract awarded to BGS
- Jun 2001 Landmark took over DataStore facility
- Sep 2003 DEAL re-launched as a data registry

Sep 2003 National Hydrocarbons Data Archive (NHDA) launched

Through DEAL, CDA provides access to released and commercial seismic and well data products relating to the offshore UK and information on oil exploration and exploitation licensing. Proprietary oil company data is made available to members-only through the CDA DataStore. Archives of data from the North Sea are maintained in the National Hydrocarbon Data Archive (NHDA) hosted by the British Geological Survey at NERC's Gilmerton site.

As noted above the CDA provides access to two data services:

1) Digital Energy Atlas & Library (DEAL)

DEAL is owned by CDA and is operated by the BGS (British Geological Survey). This data registry is a web-based service, designed to promote and facilitate access to data and information relevant to the exploration and production of hydrocarbons on the United Kingdom continental shelf (UKCS). DEAL is therefore mainly a metadata store.

DEAL is developing into a full national catalogue of UKCS geoscience data, and through DEAL's web-based Geographical Information System (GIS) the Data Registry points users to the source of selected data of interest, or users can access networked repositories of UKCS geo-scientific data as a single unified data resource. It is available for the public via the Internet, but access controls are imposed on CDA well data.

DEAL is a source of spatial and attribute data for the UKCS. The following types of data are available:

- Seismic survey records
- Well header records
- Licence and block geometry (marking out areas on the sea bed)
- Pipeline alignments
- Platform locations
- Oil and Gas Field outlines
- Coastlines
- International boundaries

DEAL obtains product metadata from catalogues supplied by vendors (oil companies), released data from companies, now in the public domain, and from data supplied to the DTI. This information is not thoroughly checked when received, except to verify: a) that it does indeed relate to the well it purports to describe, b) that its structure is correct.

The registry points users to the various data resources held at third parties: the NHDA, data stored within member companies data bases, the BGS, and elsewhere, as well as to the CDA DataStore. Depending on the resource accessed, data is available on-line or can be downloaded.

2) CDA DataStore

The services for well data are provided under contract by Landmark EAME Limited. From July 2001 the CDA DataStore could also be accessed via DEAL.

The CDA DataStore originally also maintained a database of 2D and 3D seismic navigation data but it is intended to this data to DEAL in the near future. The data is available to those companies that are participants in what is called the CDA's "Seismic Phase". By agreement between the CDA participants, all seismic navigation data submitted to the repository is available to all participants.

The following data are available:

Well data:-

- Raw digital well logs (from participants)
- Joined digital well logs (from participants)
- Scanned well reports (from participants)
- Scanned well logs (from participants)
- Scanned well reports & logs
- Well deviation data

Seismic navigation data (i.e. from traverses of the sea bed to obtain seismic information):-

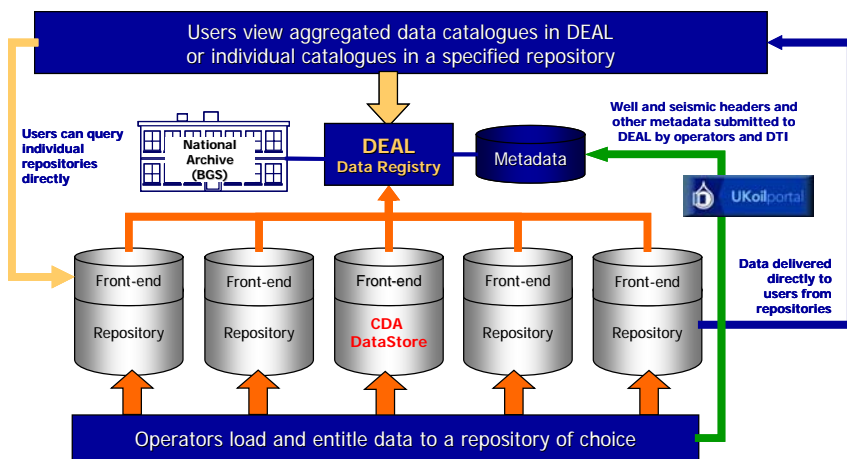
- 2D seismic navigation data
- 3D seismic navigation data
- Sailing-line data
- Grid geometry data

Some **public data** is also available:-

- Exploration licence data and licence history
- Well headers (locations and other identifying attributes)

As at October 2004 there were 175,000 scanned images, 4,250 seismic navigations, and some 7,000,000 "curves" (profiles down the well borehole) from about 9,000 drilled wells. Legacy data has been supplied by CDA participants. The DTI's 2D seismic navigation database was adopted and loaded to provide a basic set of pre-1993 2D seismic data.

The diagram shown below (courtesy of CDA Limited) shows the basic data access and delivery routes.

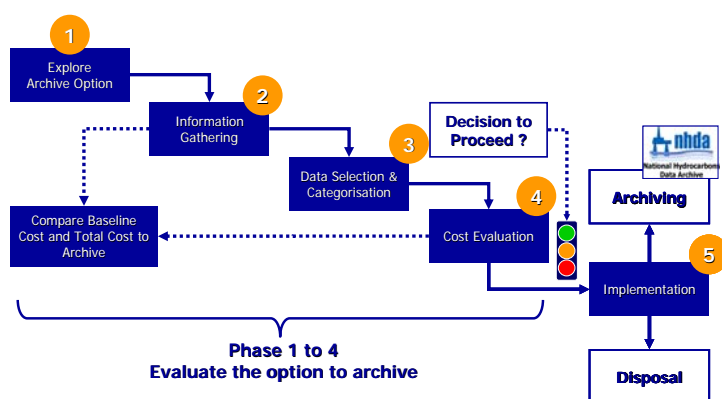


Support functions for users are available according to the different data resources accessed. Help is available from CDA itself, which also hosts a well designed and informative web site (including a members’ private area); help can also be obtained from the BGS and NHDA, and CDA DataStore.

Archiving

The CDA and associates have a sophisticated approach to archiving data. Those who hold licences to explore and exploit the UKCS are required to retain licence data in an accurate and usable form in perpetuity (this obligation survives data release and licence relinquishment).

The CDA maintains that pooled archiving makes considerable sense for its members, helping them make cost savings on data storage and management, and to undertake data preservation and so that duplicated and redundant data can be eliminated. An archiving Handbook is available at <http://www.bgs.ac.uk/NHDA>; and a five-step process is proposed to evaluate, on a cost payback basis, whether data should be archived, disposed of (if permitted) or kept active. These steps are summarised in the following diagram.



Operational issues

Landmark EAME, who run the CDA's data storage systems, also perform data conversions (e.g. scanning), data loading, data management. The company employs approximately 10 people on this..

The CDA are also working on providing pipeline data; this includes information on pipe capacities so that customers can assess the potential to add further production volume on them.

The CDA provides a gateway to Terabytes of data. Much of this is still on millions of magnetic tapes.

Standards

Geophysical data are well defined, as everyone needs to acquire, store and access the same information. This was achieved by the oil industry getting together, and seeing commercial benefits in doing this. The effort went across all data types at the raw data level. But it did not extend at all to metadata, and nomenclature still remains a problem.

For metadata standards and nomenclature (e.g. company names, well names) there are too many standards, causing problems with obtaining a common understanding.. The DTI has one, but it is old and based on 80-column cards. Mergers and acquisitions also cause problems, including merging of catalogues.

There are also difficulties at the interface between what our informants called the scientific level and well level. This is illustrated by changes with time and the turnover of technologies used in exploration, leading to changes in terminology and semantics.

Well data formats "work". However, practice has thrown up a lot of problems, and it is clear that one needs to understand the data to use it.

The CDA does not want to be a standards body. They did develop a few standards: CS8 (a high level standard), CS9 (seismic standard), CS3 (naming). They developed them because they did not exist. However, they would rather manage this activity, not undertake it.

Standardisation processes

CDA standards will follow industry initiatives and trends and will reflect UK legislative standards as defined by the DTI. CDA strongly encourages participation from both service and oil companies in the construction and development of CDA standards and practices.

CDA has a Standards sub-committee. This group, formed of some six individuals from CDA members, considers each individual standard and redrafts, adopts or drops standards as necessary to meet the changing needs of the business. The sub-committee takes into account de facto standards that are in use, practices in similar initiatives, and practices in other environments. The sub-committee produces a draft standard for submission to the CDA Council which decides whether the standard is to be mandatory or a recommended practice.

Views on data sharing

The CDA noted that use of the system is increasing, though described their user group as generally “somewhat passive” (rather than pro-active). A survey was made of DEAL users, asking what use they make of it and why, and inviting comments on the user interface design. It emerged that most users have a very focused, narrow view of what they want; they typically have one or two specific tasks to perform, and access the data resources just to answer specific questions related to these. Users ranged from technophobes to technophiles.

It was noted that as a whole geophysicists tend to use maps, not forms, and tend to work very much in an ad hoc fashion. This suits the discipline.

The CDA believed the main characteristics driving their success are the quality of the data, and its currency and completeness.

Regarding data quality, it was noted they had a “verified” flag attached to data, but this has not proved particularly useful (compliance with assigning the flag could be a problem). They now have audit methods, noting that “history of acceptance of the data is their best guide”.

Reference web-sites:

CDA:	http://www.cdal.com/
DEAL:	http://www.ukdeal.co.uk/
UKOOA:	http://www.oilandgas.org.uk/
PILOT:	http://www.pilottaskforce.co.uk/
National Hydrocarbons Data Archive (NHDA):	http://www.ac.uk/NHDA/home.html

Addendum 1: CDA Members at October 2004

Amerada Hess	Newfield Exploration
BG Group	Paladin Resources
BP plc	Perenco
Burlington Resources	Petro-Canada
Caledonia Oil and Gas	Samson Int'l
ChevronTexaco	Shell U.K. E&P
CNR International	Statoil
ConocoPhillips	Talisman Energy
DONG	Total E&P
EnCana	Venture Production
ENI	
EOG Resources	Aceca
ExxonMobil	DTI
Kerr-McGee	Edinburgh University
Marathon Oil	Petrotechnical Open Standards Consortium (POSC)
Murphy Petroleum	

Appendix 3.6 Ensembl

Key data**Name:** **Ensembl****Acronym:** **Ensembl****Established:** 2000

Brief description: Ensembl is a joint project between the EMBL-EBI and the Wellcome Trust Sanger Institute that aims at developing a system that maintains automatic annotation of large eukaryotic genomes. Access to all the software and data is free and without constraints of any kind. The project is primarily funded by the Wellcome Trust. It is a comprehensive source of stable annotation with confirmed gene predictions that have been integrated from external data sources. Ensembl annotates known genes and predicts new ones, with functional annotation from InterPro, OMIM, SAGE and gene families.

Contact: Ewan Birney

Wellcome Trust Genome Campus
 Hinxton
 Cambridge
 CB10 1SD

See: <http://www.ensembl.org/index.html>

Staffing: ca. 35 staff, shared over the Sanger Centre and the EBI

Interview data: Ewan Birney (Senior Scientist, EMBL-EBI) by Alison Macdonald and Philip Lord, 6th September 2004, and 11th January 2005, with Richard Durbin*, Tim Hubbard*, Arek Kasprzyk, Steve Searle*, Arne Stabenau, James Stalker* (* Wellcome Trust Sanger Institute, all others EMBL EBI)

Ewan Birney and six his managers, and held a seminar with us at Hinxton during one of these meetings; Tim and Richard were present for the first 30 min.

Ensembl Case study report

Overview of resource

Ensembl is a joint project between the Wellcome Trust Sanger Institute and EMBL-EBI that aims at developing a system that maintains automatic annotation of large eukaryotic genomes. Access to all the software and data is free and without constraints of any kind. The project is primarily funded by the Wellcome Trust. It is a comprehensive source of stable annotation with confirmed gene predictions that have been integrated from external data sources. Ensembl annotates known genes and predicts new ones, with functional annotation from InterPro, OMIM, SAGE and gene families. The Ensembl project offers an integrated, extensible, and re-usable framework for generating, storing, retrieving, and displaying genomic annotation data, data whose value and importance to the scientific and research community will continue over decades and centuries.

Currently information is available on 17 species (see list below), including human, dog, chimp, rat, mouse and domestic fowl.

Mammals

- *Bos taurus*
- *Canis familiaris*
- *Homo sapiens*
- *Mus musculus*
- *Pan troglodytes*
- *Rattus norvegicus*

Other chordates

- *Ciona intestinalis*
- *Danio rerio*
- *Fugu rubripes*
- *Gallus gallus*
- *Tetraodon nigroviridis*
- *Xenopus tropicalis*

Other eukaryotes

- *Anopheles gambiae*
- *Apis mellifera*
- *Caenorhabditis elegans*
- *Drosophila melanogaster*
- *Saccharomyces cerevisiae*

The Ensembl web site is excellent (see: <http://www.ensembl.org/>) and acts as the principal user interface to the data. It is rich in functionality (for detail, see [15] and [16]). The site currently serves over 500,000 pages from 2.5 million hits per week, providing access to more than 80 GB (gigabyte) of data to users in more than 80 countries. Ensembl caters for a global community of users who vary considerably in their needs and demands.

The user interface is home grown. The database is built on top of an open-source platform comprising Apache/mod_perl and the MySQL relational database management system. The software and database are modular, extensible. There is a lot of heuristics (the application of knowledge derived from experience) built into the program code. The software is actively re-used and extended in many different projects, and has been downloaded and installed in companies and academic institutions worldwide.

Both software and data are freely available.

There is a help desk available online and via telephone. The Ensembl Web site maintains user preferences (such as which tracks are displayed in ContigView) through the use of cookies.

Scope of resource

The question of depth vs breadth in the database was an active debate a little over a year ago, but was resolved by their advisory board: they should focus on vertebrates only, but establish “connectors” to other orders by including some non-vertebrate species. This was a strategic decision; Ensembl has a fundamentally medical, mammalian remit. The work on *Arabidopsis* at Nottingham where they have adopted the Ensembl toolset is a good example of how this connector approach worked, in this case with a plant species (see the NASC case study). Ensembl has a scientific advisory board, which the project team values and to which it can turn for guidance on strategic decisions.

Regarding specificity vs sensitivity of the searches and annotations they were forced to make a decision on this and deliberately went for specificity. Although users initially criticised this, they have since been vindicated in their stance. Finding false positives is a waste of time and money for consumers.

Staffing and recruitment

There is a staff of some 35 people working on Ensembl, shared across the Sanger Institute and the EBI, at the Genome Campus in Hinxton, employed either by the Sanger Institute or the EBI. Team loyalties to the project are very strong (and there is strong competition to join). The team has a very flat organisational structure, with small, fairly informal project groups of two to six people. By having a different co-ordinator each month for the monthly data release, project team members gain experience and insight, and can make more informed contributions to the project. An additional benefit of the system is that the project is substantially less exposed to risk related to the loss of personnel.

Among the managers who were interviewed, recruitment policies differ. The database group take programmers, and teach them the biology they need. The gene-build group take biologists with some computing interest and train them up to the computing tasks. The Sanger Institute provides good training. Participation in scientific meetings is also seen to be an important part of training.

Interviewees felt that MSc courses in bioinformatics are not rigorous enough, nor practical enough, finding that these graduates have very little experience of writing large software programs. People who have had an outplacement with pharmaceutical companies are a useful source of candidates. (The Sanger would also like to take on such placements, but do not have the resources to undertake it.)

Sanger has formal training courses, EBI does not (for lack of funding). Ensembl needs internal training for new recruits, extending over two to three months as there is a large program code base. Programmers need to have understanding of genome biology, and they do try to recruit programmers with this background.

Operational aspects

A key informant stressed the importance of database planning. Ensembl works with a forward medium-term horizon of three to four years, for robustness. Database management, he noted, has three aspects: content, function, and management, the first two driving the

latter. Our interviewee remarked that the complexity of running the Ensembl database extends far beyond just “the way the data works”, and noted the similarities with banks: there is a balance between how much they can develop and the robustness of the database, and they have very similar toolsets.

Operational planning works to a six-month horizon, which gives good control. Ensembl also tests potential projects by using pilot projects.

We explored the monthly release cycle for data and software the team provide. They noted that from an academic perspective the process is very tightly controlled and highly structured; however it is much less rigid than a similar commercial operation would be (for example there is less cross-checking of code before release). They have to maintain a balance between commercial practice and academic looseness, and between speed and quality. The user community to some extent acts as code reviewers (but there are different levels of user – some merely use the web interface, while others download the whole system).

The cycle is driven by a number of key deadlines during each month: thus on the 1st of the month the genome to be released is decided (updating is done by genome); on the 6th data is handed over for pre-processing; and so on through the month in a series of defined steps. Each month there is a different release coordinator, rotating through the managers. The coordinator keeps track of progress, double checks the release (“health checks”), takes any tough decisions if problems arise. As the release date approaches, data and program code fixes, and web site updates are brought together as a package; the whole system is rebuilt and release takes place on the first of the next month, and the cycle repeats.

They stated that it was very good to have a fixed date on which to bring data, code together – it focused effort. In driving this cycle, data is more important than code, and it is more important to get that right – it is the data that most people really want.

Occasionally they make “revolutionary” updates, when major changes are implemented. There have been only been two or three of these.

Many users are unaware that the update cycle takes place. For some however, monthly release cycle is very punishing (for example for those who write extensions to the code), so some tend to take code at periodicity which suits them such as annually, quarterly.

Adding a new genome is easy, since the update process always starts from scratch. When asked about the costs of adding a new genome they said it was very easy. Compute costs are very roughly proportional to n^2 , where n is the number of genomes (species) – there is a combinatorial increase with each new genome.

They have a yearly turnover of computing equipment to cope with upgrades, obsolescence and expansion.

Success factors

Ensembl is one of the major resources based at the Hinxton Genome Campus, working with and as part of the fabric of resources (data and tools) available from the Sanger Institute and the EMBL-EBI, where data and tools are consciously brought together, with inter-operability one of the primary objectives. An interviewee noted that the explicit separation between

services and research at the EBI, noting the “creativity aspect” to this infrastructure, with research blending into development, into the production process.

One of the factors which has contributed to their success is very strong collaboration with the systems group at Hinxton, which runs a 1000-node computer farm in which the Ensembl infrastructure runs: Ensembl has “incredibly good feedback from their systems group”. This group has people with a science background, which helps them understand Ensembl’s needs.

The Ensembl user interface and tools are easy to use, and so the data and tools are very widely used. The assumption is often that the tools and interface were easy to build, but behind this ease of use, however, lies enormous work and skill, making the tools web-enabled, for example, and “making the difficult bits disappear”. The risk to such endeavours is that funders also think that little work is required to achieve ease of use, making it difficult for such projects to argue for funding for more than a few people.

They have created a virtuous circle: they have money to continue to provide a good service; good service is a consequence of the database and software being made freely available, and being of high quality; this leads to have a high scientific impact; and in turn this recognition satisfies the funders’ motives of scientific benefit and scientific recognition which thus promotes funding.

Another success factor is the very good feedback they get from the user community. However, there is no formal user group, but they do maintain mailing lists, though these are skewed away from the average user. The wide and deep take-up is also buoyed by the discoverability and outreach achieved by the project, built on design work, both in look-and-feel and in technical and scientific design.

The view was taken that making all information available (data and software) and stable and adequate funding (not only initially, but importantly on a **continuing** basis over the medium term) both contributed to success.

One suggestion for sustaining quality of provision over the medium term was to put provision of the project up for tender every five years, initiating the tender after four years and allowing other groups to apply for pilot grants, “to allow them to mount a serious bid”.

It was also noted that friendly competitive tensions **within** the team rather than with external groups was valuable. (The informants’ view being that competition raises quality; an example given was the development of the Ensembl APIs, application programme interfaces).

A point agreed by several interviewees was that they had been helped by the fact that the Human Genome Project had engendered and fostered a spirit of open-ness in people.

Data sharing

We note here some the views expressed on data sharing.

It was noted that if you are secretive in what you are doing you become a threat to “competitors”. They felt they were probably the most open of the groups working in this area (but the University of California at Santa Cruz is also very open in its operations as is the National Center for Biotechnology Information, NCBI, in the USA). They are occasionally

approached by people who want to “collaborate in a closed way” – they refuse such approaches as it would be destructive of the sharing culture.

It was noted that the major driver in science is new data, leading to publication. Science has changed so there are now many scientists, and there is a high chance that more than one person has the same idea – so concealment is counterproductive to gaining recognition. This drives early release of information.

Another point made was that the USA believes that open funding policies further commercial exploitation and trade. It has been shown this works. On the other hand, the UK and Europe tend to take a more restrictive and exclusive approach, so as to take research in a direction which it is hoped will lead directly to technology transfer and commercial exploitation. This model has been shown to work less well in this regard.

Cost, benefits and risks

With SNP data there is no way of knowing if it is right or wrong – it is generated algorithmically. However, many millions of pounds’ worth of others’ experiments rely on the data. This is a heavy responsibility on the skills of the group. Ewan Birney said that they had saved the cost of Ensembl many times over in saved expenditure on reagents by users.

Frequently heard in the meeting was warm appreciation for the Wellcome Trust model – open access and their type of funding.

Appendix 3.7 Genome Information Management System

Key data

Name:	Genome Information Management System
Acronym:	GIMS
Started:	1999
Brief description:	The Genome Information Management System (GIMS) is a data store system which focuses on analysis of data. It integrates genome sequence data with functional data on transcriptome and protein-protein interactions in a single data warehouse. GIMS stores the data in a way that reflects the underlying mechanisms in the organism (in this case, the <i>Saccharomyces cerevisiae</i> (yeast) genome), permits complex analyses to be performed on the data, and provides users with pre-written queries.
Contact:	Dr. Michael Cornell Department of Computer Science, The University of Manchester Oxford Road Manchester M13 9PL GIMS Project: http://www.cs.man.ac.uk/img/gims/details.html
Funding:	GIMS was funded by the BBSRC/EPSRC Bioinformatics programme, the Wellcome Trust and by the COGEME Consortium for the Functional Genomics of Microbial Eukaryotes.
Staffing:	Professors Norman Paton, Stephen Oliver, Andy Brass; Research Staff: Dr. Michael Cornell, Cornelia Hedeler
Data held:	Genome sequence data and functional data on the transcriptome and protein-protein interactions, relating to the yeast genome, <i>Saccharomyces cerevisiae</i> .
Interview data:	Professor Stephen Oliver on 28 October 2004; Dr. Michael Cornell, Michael Williamson, Kevin Garwood on 28 October 2004; interviews by Denise Ecklund and Alison Macdonald.

GIMS Case study report

GIMS is an interesting case study for data sharing for several reasons: GIMS is a tool for holding and sharing different types of data, queries and analyses, but project team members are also users of data resources; GIMS addresses functional genomic questions involving multiple, genome-level data sets, and it is principal-investigator led.

Underlying the GIMS project is the hypothesis that a full understanding of the function of a gene requires different data sets obtained at different levels of genome-wide analysis to be integrated. GIMS develops an environment specifically to support integrated analysis of multiple types of genomic data. This capability was not available at the major genomic resources such as MIPS (the Munich Information Centre for Protein Sequences), KEGG (the Kyoto Encyclopaedia of Genes and Genomes), or YPD (the Yeast Proteome Database). These support browsing and visualization, rather than analysis.

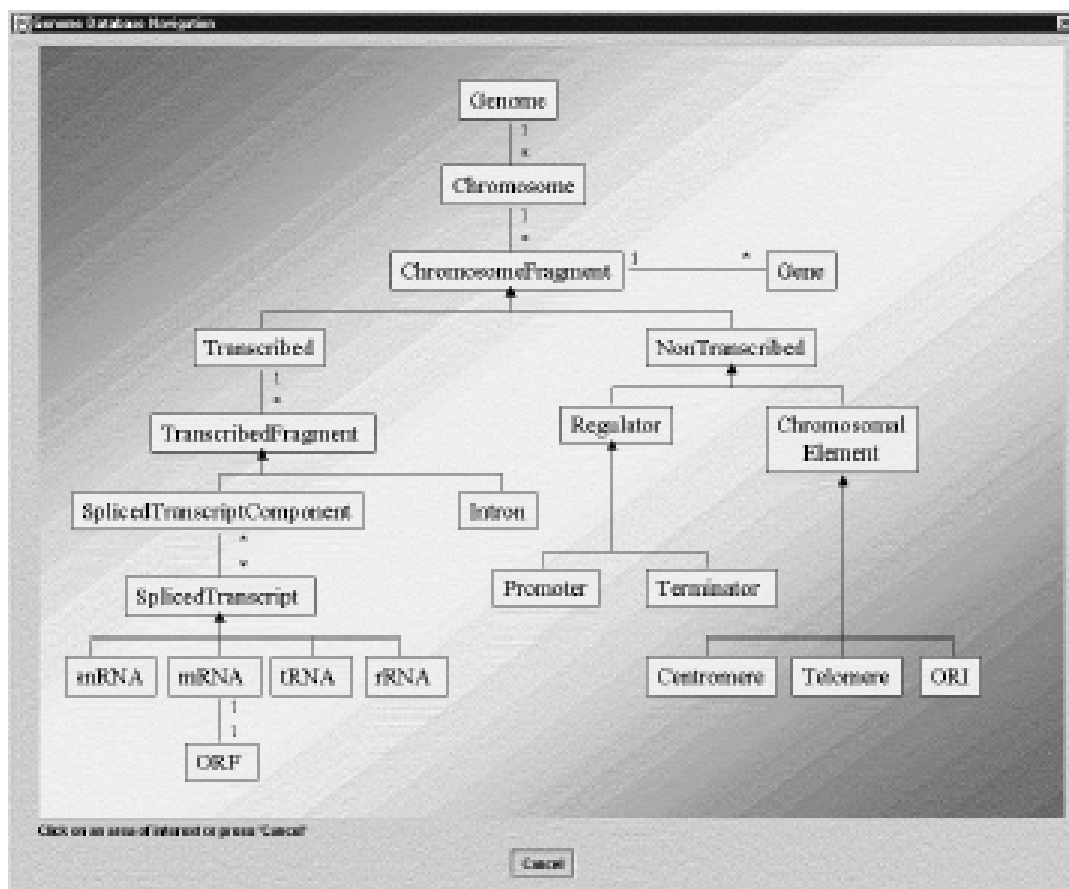


Figure 1: The GIMS browser showing the genome schema at class level. Rectangles represent classes, lines show the relationships between classes, and numbers and “*” show the number of objects participating in the relationship. Arrows point from sub-class to super-class. Clicking on one of the classes executes the GIMS browsing programme.

Source: ([29]Cornell et al., IEEE 2001

To support complex and rich analysis of different data types, the GIMS system used the “object” database form, which models and holds data as objects. This means that a database can mirror a conceptual model of data, which in turn allows rich and “more intuitive representations of stored data for browsing and programming” [Cornell, 2001]. The GIMS

system has two principal components, a data model and a “canned” query interface. Using Unified Modeling Language (UML), the database schema represents the genome sequence, protein-protein interactions, mutation and phenotype data, and the transcriptome (fully described in Paton, Khan et al., *Conceptual modelling of genomic information, Bioinformatics*, 16(6), 2000). This object model has been carefully designed, replicating sequence and function data. This environment means that it is a straightforward matter to construct queries and analyses, which in other database environments would be difficult to express.

A “canned” query is one which has been pre-written, and resides in the system. The GIMS system provides an interface through which users can run these pre-programmed queries: a canned query is an analysis task with parameters. Clicking on a listed query brings up a form which requests the values for the user to provide, giving the parameters of the query. The GIMS infrastructure is three-tier, so that this query interface can be run as a small application (applet) which can be sent along the Web to a user or as a stand-alone program.

The analyses enabled, which worked across multiple data sets, showed that obtaining insights into specific functional data sets were often easier when made with reference to other categories of data.

GIMS uses an inexpensive commercial object-oriented database, FAST Objects. Users do not have to purchase the FAST Objects software to use GIMS, unless they wish to set up their copy of GIMS.

GIMS database – data population

The GIMS database draws data from the *Saccharomyces Genome Database (SGD)* and *MIPS* (sequence and related data), transcriptome data from the *Stanford Microarray Database* or from the *Manchester University Yeast Group*. Other data is collected from a variety of other sources.

The number of sources is minimized to reduce possible inconsistencies in naming and numbering of sequences. These inconsistencies were raised by several key informants, in this and other contexts, as an obstacle at semantic level to easy or automated integration or comparison of different data sets. Such hurdles could often be avoided by use of standard naming conventions. Several key informants referred to tables which had had to be compiled to enable efficient translation between databases which apply different naming conventions. For a sustained translation capability, these tables need to remain available and up to date, unless the databases’ naming systems are co-ordinated.

The population of the database involves several types of action. Some data is replicated from public databases such as *SGD* and the *Stanford Microarray data*; other data is extracted from the *MIPS protein interaction database* and other published experiments, found by active searches of journal papers. Difficulties were encountered which many other key informants reported: there was no notification to the community when a database had been updated, or its schema changed. Monitoring the databases and trawling for data are very time-consuming. Another difficulty was obtaining all the metadata that was needed, which was often only partially available.

GIMS users, development environment

Users share the GIMS data and services by accessing a single copy of the GIMS software and the GIMS repository. Most users are in the School of Biological Sciences at the University of Manchester; external users access GIMS by remote log-in. Some external users have set up their own GIMS systems, using the software and database schema, with guidance from the Manchester GIMS group. The Manchester GIMS development group has also provided a Linux version, and a version for the Apple Macintosh operating system (in Australia). Both these instances are managed and run independently – the different databases are not federated, and have diverged. Another Australian group is using the GIMS software for a data warehouse for mouse genome data.

The GIMS web site provides a guide for users which is clear, easy to read and to follow. However, the number of downloads of the software for accessing the Manchester GIMS warehouse was depressed by the need to download and install a small and simple piece of software, even though a 15-minute tutorial is provided with the download package.

The need for this installation also meant that the GIMS database was not eligible for inclusion in one of the community's major lists of database resources, because it was not web-based – even though its form was the reason why the database enabled rich analyses. The GIMS database is being extended to support web-based use.

The GIMS development team combine computer science and biological expertise. Furthermore the two departments (computer science, biological sciences) are located close together, with very close interaction and many joint projects. This geographical co-location was stressed as a very important factor, similar to the proximity of the EBI and the Sanger Institute, facilitating interaction between their groups.

Many bench scientists have little computer knowledge beyond familiarity with the basics of the software packages they use – the Microsoft programs, statistical programs. Speaking more generally, the interviewees noted that for many, unless something looks like a web interface and involves installing something, however simple, they will shy away from using the tool. But the interviewees stressed that software functionality can be significantly inhibited by a web-based format.

They noted reluctance on the part of many researchers to use tools developed for them. In some cases this was because of mistrust of the digital medium – fear that electronic capture would mean their data would be shared with others before publication of their own results. In other cases the fear was that, because the tool was not accessed using a web browser it would be too difficult for them. In many cases, however, tools were not used because they involved doing something unfamiliar. However, as soon as a training course was run, not only were the tools used, but the users also provided feedback about features which would be useful.

Further points highlighted by interviewees were the benefit of good centralized management of software projects in the Manchester development group, such as for MaxdLoad (a tool for the loading and submission of microarray data), and collaboration between research groups. Collaboration between research groups means a greater level of shared knowledge and good practice in software development (such as established software development procedures, a repository for records of software code changes). The presence of a broad skills set at Manchester also means, for instance, that there is a pool of knowledgeable end users who can test any tools which are developed.

Next stages

As well as web-enabling GIMS, the development group will be developing more analysis tools and enabling the system for a Grid environment. GIMS will also be used and extended in a follow-on project, e-FUNGI. This project will integrate sequence and functional data from multiple fungal sequences in a way that facilitates the systematic study of less well understood species with reference to model organisms with more fully explored functional characteristics. It will also support comparative functional genomics through the development of libraries of bioinformatics queries and analyses that make use of the warehouse, and which can be combined in different ways to conduct studies of cellular processes, pathogenicity and evolution. See: (<http://www.e-fungi.org.uk>)

As the principal investigator noted, funding for projects is for a limited span only. Often these projects generate data, digital processes or tools whose life should extend beyond that span. In some cases, public repositories exist for data, such as the ArrayExpress resource. For other data, however, no provision is made other than unofficial and unfunded.

An interviewee also noted that GIMS and other projects based on access to integrated data would benefit from access to more databases, held by third parties. This access has not been provided hitherto because administrators or database owners were worried that access by others will compromise their data – values might be altered through handling errors, the database might even be broken. The interviewee's suggestion was that access could be provided not to the original but to surrogate copy databases.

Appendix 3.8 Malaria (*Plasmodium falciparum*)**Key data**

Name:	Malaria <i>Plasmodium falciparum</i> genome
Acronym:	Malaria
Started:	1996
Brief description:	The malaria parasite's (<i>Plasmodium falciparum</i>) genome was sequenced during the late 1990s and is now available in curated form at GeneDB at the Sanger Institute at Hinxton, and PlasmoDB at the University of Pennsylvania. This information is made available for free, and is aiding the search for drugs and vaccines to combat malaria.
Contact:	Dr. Andrew Berry The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA http://www.genedb.org/ is the Sanger Institute Pathogen Sequencing unit home page http://plasmodb.org/ is the Plasmodium falciparum database resource housed at the University of Pennsylvania, USA
Funding:	Sequencing project: £18 million, funded by the Wellcome Trust, The Burroughs Wellcome Fund, the National Institute of Allergy and Infectious Diseases, and the US Department of Defense.
Staffing:	Dr. Andrew Berry is curator of the Sanger Institute <i>Plasmodium falciparum</i> database
Data held:	At Plasmo DB: <i>Plasmodium falciparum</i> sequencing data: the database integrates the DNA sequence with automated and curated annotation, RNA and protein expression data, cross-genome comparisons, and tools and emerging genomic and proteomic-scale datasets. At GeneDB, a multi-organism database holding data produced by the Wellcome Trust Sanger Institute and other sequencing centres: manually curated <i>Plasmodium falciparum</i> sequencing data, within the pathogen sequencing resource; this data is updated weekly.
Interview data:	Professor Chris Newbold on 17 th December 2004; Dr. Andrew Berry on 22 nd December 2004; both interviews by Philip Lord and Alison Macdonald.

Malaria Case study report

The malaria genome sequencing project and resource are of great interest from a data-sharing perspective on many counts, but above all for the role of individual research community members and funders in instigating the project and taking it through to its successful release as a resource freely and immediately open to all, world-wide, and for the continued availability and development of the resource. Exemplary also are the project's involvement of the global malaria research community throughout the project pre- and post-sequencing, the weight placed on sharing of information and resources, and reflected in the malaria project consortium's open-data policy, its use of existing tools and expertise, and its own development of tools for the malaria research community.

Genesis of the resource

In the mid 1990s, malaria researchers working on individual malaria chromosome sequencing in separate initiatives in the UK and Australia heard about each other, and that a project was starting in the USA. These researchers met, approached funders, and a consortium project to sequence the genome of *Plasmodium falciparum* (the most virulent of the four species of malaria parasite) was born. With an "AT" content of some 80%, this genome presented a particularly challenging sequencing endeavour, including at computing level, requiring computer-alignment algorithms to be rewritten and novel gene-model software to be developed.

Without the engagement and commitment of the funders, neither the project nor the resource would have been possible. Throughout the sequencing project, the Wellcome Trust provided wider support, hosting meetings not just on the specific project but driving other similar resources and open-access policies and declarations. As an interviewee noted, "the funders are the people who can make the difference". He also noted, "there is a potent argument that big data sets should come out very fast": funding enables that speed, bringing the resource rapidly to a large number of researchers and laboratories. The funders not only supported the sequencing, but also the immediate availability and continued development of the resource, its hosting, in addition to other malaria resources such as the malaria reagents resource. This resource is consulted and used daily by tens of thousands of researchers from every country in the world.

Individual researchers had seen how swiftly basic biology could advance when information from genome projects is used by a wide variety of researchers, supporting their own fundamental ethos of open access to information and data.

From the outset, researchers and funding agencies actively pursued the objective of making the data available in as accessible, non-exclusive, rich and practical way as possible to users, wherever they may be. Regular and carefully planned meetings and workshops were held, over the different continents of the consortium, which specifically examined how to improve sharing of information and resources. From the very early stages, the project was thinking forward about ways to capitalize on the genome information.

By using existing resources and expertise, the project not only avoided having to "reinvent the wheel" but thereby made the resource more inter-operable with other resources. Meeting records reflect the project's active efforts to call upon other expertise and relevant techniques, with speakers such as Professor Michael Ashburner on the FLY database originally set up for

the *Drosophila* research community but more widely used and linked to other organism databases, Professor Steven Oliver on the yeast sequencing project, Martin Aslett on curation of other genomes. The latter stressed the need for analysis to be maintained once sequencing had been completed, with a dedicated person working early on the project.

Special emphasis was made throughout the project and subsequently to involve the global malaria research community, including those in “endemic” countries, and to enable electronic access to malaria genome sequence and functionality. This was reflected at publication, with the insistence that a CD-ROM of the sequence for researchers with poor Internet access be included in the Nature publication.

With a project consortium spread over several continents and research groups, agreeing a publication date was important. The project found that setting a date, agreeing that date with the publication journal (Nature, October 2002) and setting deadlines for papers substantially quickened progress. At an interviewee’s suggestion, the publishing strategy was adopted of a single, joint date for publication of the genome, with Nature running a cover story. The data release policy is that researchers could publish any papers they wish, but the only limitation on use of the data is that no single researcher or group can do a paper on the whole genome.

Collaborative development

During the project, the project leaders went to considerable and sustained effort to ensure that the whole malaria community could engage in the project and were canvassed on major issues. They invited people to early meetings, sent open emails to researchers using list servers, thereby achieving community engagement, contributions and support for decisions.

Database development

From the outset the researchers and funders wished to develop a user-friendly, practical, centralized database for analysing and using the malaria genome information. At first they feared that the database development costs would be high, potentially a barrier to the project. However, by making use of existing database work they found that these costs were a lot less than feared.

Database design, development and issues were explored, discussed, agreed and planned, always guided by the project team’s objectives of providing a resource providing fast, useful, rich access to the data. Meetings identified a wish-list of features and functionality, and the resources needed to enable these. Again, meetings were determined that the malaria research community was actively consulted throughout the development process to ensure that the resource meets their needs – indeed, the community wished to be involved.

One of the important challenges was to set up a highly robust data exchange capability across continents. This capability is still in use. Problems they had to address were differences in database schemas, different systems, and different formats. Substantial work was done to map across the different systems.

An interviewee noted that co-ordinated, effective genome curation efforts are important to maximising the investment made in the sequence, and that this effort must evolve as new data and new tools and methods become available. Curation in this context comprises a number of

functions which are designed to facilitate querying and retrieval of data. Experienced curators are employed at the Sanger Institute to curate six of the organisms in GeneDB, including *Plasmodium falciparum*. This parallel curation of genomes at the same institution and within the same environment introduces a degree of consistency in annotation. Curation of related species in the same environment also allows for richer cross referencing (e.g. across *Plasmodium* species) and to enable comparative studies to be facilitated.

Within Gene DB an effort is made to keep sequences, annotations etc. up to date as new information is published, information is contributed and new submissions are made to public databases.

GeneDB uses the Genomics Unified Schema (GUS), a relational database schema and associated application framework designed to store, integrate, analyze and present functional genomics data. The GUS schema supports a wide range of data types including genomics, gene expression, transcript assemblies and proteomics. It places emphasis on standards-based ontologies. In addition the GUS provides various tools kits (see www.gusdb.org). Use of GUS as a standard helps with the exchange of data and tools between databases – and this helps users too as the interface for searches within Plasmo DB and GeneDB will be familiar.

Use of Gene Ontology

Informants stressed the value of consistent and recognised vocabularies. Gene Ontology (GO) is recognised as a leading effort in this field. GeneDB is a member of the GO Consortium (See <http://www.geneontology.org/index.shtml>). This consortium has created an extensive set of tools as well as an ontology, which are provided free to academic users.

Appendix 3.9: Nottingham Arabidopsis Stock Centre

Key data

Name:	Nottingham Arabidopsis Stock Centre
Acronym:	NASC
Established:	1991
Brief description:	The Nottingham Arabidopsis Stock Centre (NASC) provides germ lines (seeds) and information resources to the <i>Arabidopsis thaliana</i> community and to the wider research community. The NASC Affymetrix Service processes Affymetrix gene chips on behalf of customers through the UK's GARNet programme.
Contact:	Dr Sean May Nottingham Arabidopsis Stock Centre (NASC) Plant Science Division, School of Biosciences University of Nottingham Sutton Bonnington Campus Loughborough LE12 5RD http://arabidopsis.info/
Funding:	Current total grant is £4.5M from variously the BBSRC, the University of Nottingham and the European Union.
Staffing:	18, plus varying small number of temporary technical staff. This comprises the director, customer service manager, seven people involved in seed distribution (four part time), some bio-informaticians, and two with the Array service.
Interview data:	Sean May (Director), Emma Humpheys (Customer Service Manager), were interviewed by Alison Macdonald and Philip Lord, 29 th October 2004. Members of software development team (Beatrice Schildknecht, Dr. Nick James and Dr. David Craigon) were also interviewed re programming issues.

NASC Case study report

The Nottingham Arabidopsis Stock Centre (NASC) is the European centre providing seed and information resources to the international Arabidopsis thaliana community and to the wider research community. Arabidopsis thaliana (abbreviated here to Arabidopsis) provides a good model for many plants, including brassicas and other crops, and has the advantage of a rapid turnover from seed to plant of about eight weeks. The plant has a world-wide distribution.

The NASC staff comprises a director, a Seed Distribution Group, a Bioinformatics group, a microarray services group, and a Customer Services Manager. The team is under very resourceful management, providing an excellent, considerate, practical and forward-thinking service on a tight budget.

NASC's activities are coordinated with those of the Arabidopsis Biological Resource Center (ABRC) based at Ohio State University, USA, managed by Randy Scholl. Together they provide a unified, efficient service for the research community world-wide. As stock centres they have a distribution agreement: NASC distributes to Europe and ABRC distributes to the Americas. Laboratories in other locations can establish their primary affiliation with either centre.

The centre works within the Genomic Arabidopsis Resource Network (GARNet) and PlaNet, a network of European plant databases. GARNet is a consortium of providers created as part of the BBSRC Investigating Gene Function (IGF) initiative, with the aim of ensuring that functional genomic technologies are available to a wide audience. GARNet has promoted and supported services such as transcriptome analysis, bioinformatics, metabolite profiling, proteome analysis and reverse and forward genetics for the Arabidopsis community. Within GARNet other service providers are the John Innes Genome Laboratory, the National Centre for Plant and Microbial Metabolomics at Rothamsted Research, and the Cambridge Centre for Proteomics.

PlaNet is an EU-supported initiative which aims to develop and deliver a high-level plant genome database for the systematic exploration of Arabidopsis and other plants. Partners alongside NASC are:

- Institute for Bioinformatics (IBI) / MIPS (Germany)
- Flanders Interuniversity Institute for Biotechnology (VIB)
- Génoplante-Info (France)
- John Innes Centre (UK)
- Plant Research International (PRI) (The Netherlands)
- Centro Nacional de Biotecnología, Madrid (CNB) (Spain)

The Arabidopsis community world-wide is extensive and well networked; for an illustration of this see addendum 1 below.

The resources NASC provides are:

- Data – various
- Arabidopsis seed and DNA stocks
- Services: the NASC Affymetrix Service processes Affymetrix gene chips on behalf of customers through the UK's GARNet programme.

These are described below in turn.

Data

NASC has developed a version of the Ensembl software from the Ensembl system developed at the EBI/Sanger Centre (see Ensembl case study) Using this version (AtEnsembl), data on the Arabidopsis genome is available free on-line. This database contains annotations from both The Institute of Genomic Research (TIGR) in the USA and the Munich Information Centre for Protein Sequences (MIPS) in Germany. These are direct imports. For TIGR no additional gene build is carried out, and a series of alignments are done to provide Affymetrix gene probes, UniProt hits, and other similarity data.

The NASC website links to other resources e.g. PlaNet, UK Plant Genetic Resources Group (UKPGRG), GARNet.

The resource also provides access to the catalogue of seed stocks; the web site provides an ontology browser for the catalogue.

AtEnsembl receives some 20,000 hits per day.

Stocks

The centre at Nottingham holds a large stock of Arabidopsis seeds (over 300,000 accessions representing 500,000 genotypes). An accession is characterized by a mutation or an ecotype. The stocks are stored in a large refrigerated room with home-made shelving and boxes. Each accession is stored in a film negative pocket, typically holding a few tens to a few hundreds of seeds. They currently send out some 50,000 vials per year. (N.B. Seeds from the same germ line are packed into small vials; orders for one or more vials are dispatched in robust tubes via the post.) The seed stocks are reproduced in greenhouses at the unit as required.

Seed stocks are donated from the Arabidopsis community world-wide, both from public sector groups and industry. Some commercial companies also contribute stocks. If stocks are donated then there are options for keeping the identity of these stocks confidential until publication.

Services

NASC operates a transcriptomics (microarray) service to process Affymetrix gene chips as a part of the GARNet consortium. Customers supply RNA extracted from their samples, and the NASC amplifies, labels and hybridises the samples to a GeneChip. They then return intensity values for every gene represented on the GeneChip to the customer. Customers

receive a CD-ROM or DVD containing their data from the Centre, both as supplied direct from the Affymetrix machine, and as annotated using the NASC microarray database software.

The primary method of making the data public from the Affymetrix Service is via NASCArrays, the NASC's microarray database. This combines the actual gene expression data with the annotation provided for each experiment. A CD-ROM delivery service is also provided (called AffyWatch).

All donated data that is processed by NASC is made public for free using the NASC database, but customers can ask for a confidentiality period (3, 6 or 12 months). As an incentive for open data, the length of time of confidentiality affects the priority for analysis in that the samples with lowest requested confidentiality are processed first. The data is also donated to ArrayExpress at the EBI and had been donated to the Arabidopsis Information Resource (TAIR) in the USA until they closed their array acquisition program in favour of NASC and the EBI. Although NASC specializes in Arabidopsis, they will also perform other analyses, thus for example chips have been run on human, wheat, aspergillus and yeast samples.

The user community for data, services and for seed stocks is international, representing both public and private institutions. For the stocks, there are customers in 86 countries - the chief ones are UK, followed by France, Germany, Belgium, Netherlands, Japan, China. Most customers are academic users, but there are also some commercial customers.

Evolution of the resource

NASC was established in April 1991 as part of the Plant Molecular Biology initiative of the then Agricultural and Food Research Council (AFRC).

The complete Arabidopsis genome was published in 2000 by the Arabidopsis Genome Initiative (AGI), which was an international collaboration to sequence the genome. The project was begun in 1996 and the genome sequencing was completed at the end of 2000 (see The Arabidopsis Genome Initiative, 2000, *Nature*, 408:796-815). NASC managed the Recombinant Inbred (RI) mapping service that was used as part of the framework assembly process for the genome sequence.

NASC has operated its Affymetrix Service since February 2002.

Seed stocks have escalated in volume: there were 200 accessions in 1991, 20,000 in 1999, and there are now some 300,000. It is one of the largest sets of genetically modified seed stocks in the world. The accessions of variants of *Arabidopsis thaliana*, now represent over half a million genotypes.

Standards and tools

NASC have been keen on the standardisation of terminology, and are introducing ontology tools. Sean May commented that if an institution is perceived to have authority in an area then people tend to adopt their terminology. He noted that having an international audience they have to be careful over language issues.

Regarding metadata, they strive to make this full: they capture MIAME information – (watering régime, light régime, time of day, . . .). People may want to search on many unexpected things (an example given was metal tolerance). To encourage metadata compliance they try to make their submission forms as easy to use as possible – otherwise the form will not be used, or people will provide inaccurate information when they complete the form. It was noted one had to be careful about trusting received or second-hand metadata: an example given was of a technician saying that the light level during propagation was estimated to about 10,000 lux, but recorded by the scientist as equal to 10,000 lux.

Regarding computing technologies it was noted that, for users who are essentially computer hobbyists, binary, non-human-readable forms like CORBA are not suitable; whereas “readable forms” like XML and HTML are accepted as practical.

NASC are keen to be involved in Grid projects, but were also aware that some perceive the technologies as difficult, and these people opt instead to use web services and tools such as BioMoby (an open-source project using web services – for example XML, SOAP - to generate an architecture for the discovery and distribution of biological data).

Their programmers are encouraged to use SQL and Perl, but this is not rigorously enforced so that they can use tools of their choice if they are appropriate. Therefore, at this level, only weak and pragmatic standards are imposed.

At a database level they have moved to SQL, specifically to MySQL for their order processing systems. The Affymetrix system is based on DB2 – noting Oracle was priced too high for them to acquire.

Reflecting the Centre’s service nature, they also put the emphasis on using established, reliable tools and techniques. This does not mean that relevant novel tools and techniques are not used, but their use for the sake of it (a “use-the-latest-toys mentality”) is discouraged.

Economic factors

As noted, data is provided free. Charging for seeds stocks is the same in the UK and USA, noting that exchange rate movements can cause problems. NASC endeavour to keep their tariffs as low possible so that no one is disadvantaged.

The seeds are disseminated for a small charge (a fraction of the cost) to discourage wastage of seeds, plus a small handling fee per order. In cases of extreme hardship (particularly outside Europe) the Centre will consider waiving the fees.

Dr May thought that “a British Library” funding model might also be appropriate for their work. There are other considerations, along the lines of not allowing them to rest on their laurels – met perhaps by introducing a commercial element or using regular re-funding applications as an incentive.

Dr May noted that NASC does (he believes to the same or in a few cases higher standard) what it takes three centres to do in the USA, despite each US operation being more heavily funded than they are.

Staffing issues

NASC's priority of maintaining an open, sharing culture comes up against pressures for a higher output of publications - which of course, in an academic context, is also traditionally the measure for career advancement and also one of the measures by which institutions are measured. NASC provides a community resource of excellence, essentially a science-driven service, where the number of papers published bears no relation to the quality of scientific and research service provided; but, as pointed out in the BADC study, the lack of career path and professional recognition for service excellence must be a factor for attracting and retaining staff.

As with many services, there is a fair amount of routine work. However, the Centre also offers excellent opportunities for people such as qualified second-career women who have skills and knowledge and wish to put them to use in interesting and steady employment.

Training is part of their role.

Sharing policies

They are open about everything, except revealing who has ordered which strains of seed (and this is under review). Customers – donors and recipients - have confidence since they know there is no threat to them. They believe that it is best to be altruistic, and find that this also fosters altruistic behaviour in return.

At the outset, there were a small number of key individuals who initiated the open, data sharing culture in this community. The community realised that having information flows is mutually beneficial to people, and saves time. They are less concerned with IP issues than perhaps other groups. Gaining institutional support was helped by there being some very percipient people in both the NSF and BBSRC (with acknowledgement Alf Game). The NSF is led by a strong desire to make the open model work.

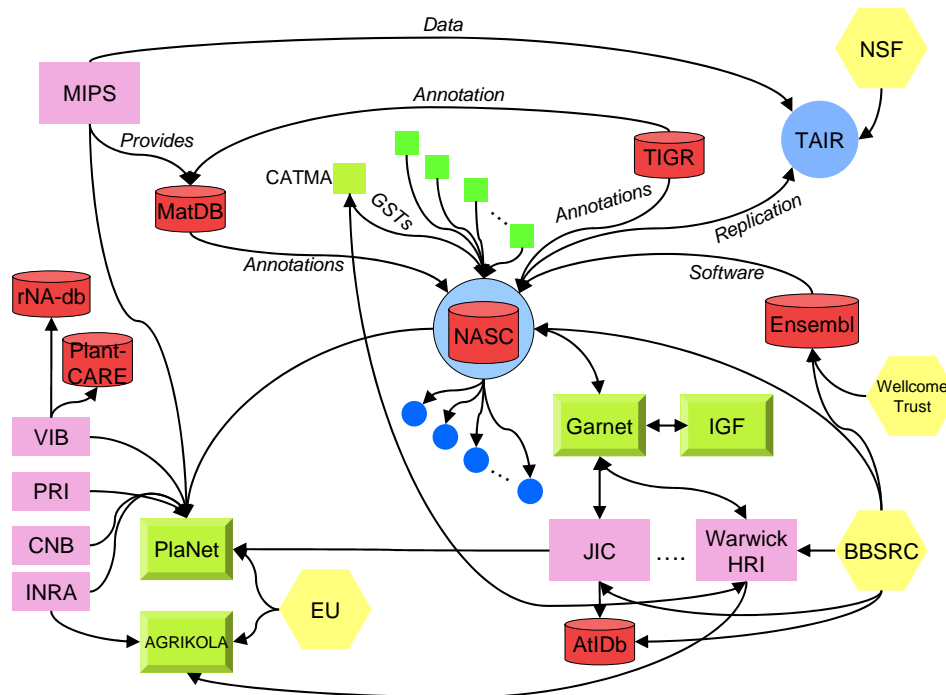
The fragility of the climate of trust and sharing was noted – indeed, it could be broken by the selfish action of a single individual.

The NASC operates between three domains: being a service organisation, a commercial entity, and an academic one. The three interfaces between these can be take up substantial time.

There is no inherent commercial value in *Arabidopsis thaliana* unlike rice, for instance, where there is and commercial problems intervene – “IP causes damage” to sharing and therefore to using rice as a plant model.

Addendum 1: The wider context of NASC

The following diagram illustrates, using the example of NASC, the complexity of the relationships entailed in data-sharing resources.



A list of the acronyms is given overleaf.

Data sharing in many of the community resource-based examples takes place within a complex web of interrelationships: with other resources, joint projects and alliances, various funders, research units. They are distributed internationally. To give a flavour of these connections we have drawn these as a diagram shown above for NASC. Though incomplete, it gives an impression of the complexity which exists. The NASC is shown in the centre as a large blue disk. Suppliers of information and data consumers are shown as green squares and blue circles. They are distributed world-wide and are numbered in thousands – individual researchers to larger laboratories.

NASC receives data from other community resources – such as annotations from MatDB managed by MIPS in Munich, and TIGR in the USA, a not-for-profit research organization. NASC developed a plant version of the ENSEMBL software with the Ensembl team. The NASC data is replicated at TAIR in the USA.

NASC is also involved in two community projects, GARNet in the UK and PlaNet, a European-wide initiative. GARNet is part of a wider Investigating Gene Function (IGF) initiative funded by BBSRC. There are a number of members of GARNet alongside NASC, including the John Innes Centre (JIC) and Warwick HRI. Alongside NASC in PlaNet are the JIC, the VIB (Belgium), PRI (Netherlands), CNB (Spain), and INRA (France). INRA is also partner in AGRIKOLA, as is Warwick HRI. Warwick HRI is involved in the CATMA initiative which supplies Gene Sequence Tags (GSTs) to NASC.

Both JIC and VIB manage further community resources in this area – AtIDb (JIC), rNA-db and Plant-CARE (VIB).

Lastly there is a well-resourced portal listing Arabidopsis information resources at TAIR which references NASC. The NASC web site also links to other resources.

Funding agencies sit in this picture – those shown are the BBSRC (funding NASC, JIC, specifically for AtIDb, and IGF, and therefore GARNet). The BBSRC are a co-funder of ENSEMBL with the Wellcome Trust. TAIR is funded by the USA's National Science Foundation; AGIKOLA and PlaNet are EU-funded projects. Other funders are omitted from the diagram, including the University of Nottingham for NASC. Ultimately the bulk of the direct funding for this whole web of activity comes from UK, European (EU and national) and US taxpayers and from charitable institutions. There is of course underlying support from infrastructures, such as the internet, W3C, and users' home institutions.

Acronyms:

Community resources / databases:

<i>AtIDb</i>	<i>Arabidopsis thaliana Insertion Database</i>
<i>MatDB</i>	<i>MIPS Arabidopsis thaliana Database</i>
<i>NASC</i>	<i>Nottingham Arabidopsis Stock Centre</i>
<i>Plant-CARE</i>	<i>European plant promoter database</i>
<i>rNA-db</i>	<i>European ribosomal RNA database</i>
<i>TIGR</i>	<i>The Institute for Genomic Research database</i>

Projects, joint initiatives, programmes:

	<i>AGRIKOLA</i>	<i>Arabidopsis Genomic RNAi Knock-out Line Analysis</i>
<i>CATMA</i>		<i>Complete Arabidopsis Transcriptome MicroArray</i>
<i>GARNet</i>		<i>Genomic Arabidopsis Resource Network</i>
<i>IGF</i>		<i>Investigating Gene Function</i>
<i>PlaNet</i>		<i>Network of European Plant Databases</i>

Organisations:

<i>CNB</i>	<i>Centro Nacional de Biotecnologia</i>	
<i>INRA</i>	<i>L'institut National de la Recherche Agronomique</i>	
<i>JIC</i>	<i>John Innes Centre</i>	
<i>MIPS</i>	<i>Munich Information centre for Protein Sequences</i>	
<i>PRI</i>	<i>Plant Research International</i>	
<i>TAIR</i>	<i>The Arabidopsis Information Resource</i>	
<i>TIGR</i>	<i>The Institute for Genomic Research</i>	
	<i>VIB</i>	<i>Vlaams Interuniversitair Instituut voor Biotechnologie</i>
<i>Warwick HRI</i>	<i>Warwick Horticulture Research International</i>	

Appendix 3.10 Proteomics Standards Initiative

Key data**Name:** Proteomics Standards Initiative**Acronym:** PSI**Established:** From 2002

Brief description: The Human Proteome Organisation (HUPO) Proteomics Standards Initiative (PSI) defines community standards for data representation in proteomics to facilitate data comparison, exchange and verification. PSI currently manages standards for two key areas of proteomics: mass spectrometry and protein-protein interaction data, and standardised formats for complete proteome pipelines are also under development. HUPO's mission is to engage in scientific and educational activities to encourage the spread of proteomics technologies and to disseminate knowledge pertaining to the human proteome and that of model organisms.

Contact: See: <http://psidev.sourceforge.net/>

Report data: Dr. Andrew Jones, BRC, including interview with Dr. Chris Taylor of the EBI.

References in the following case study are also given at the end of the case study, and the numbering refers to these, rather than to the overall bibliography.

PSI Case study report

1 Introduction

In the last decade, a new methodology has arisen to measure the localisation, abundance and interactions of proteins on a large scale, known as proteomics. The large-scale analysis of proteins has only become possible due to the success of the genome sequencing projects for all the model organisms. Many of the techniques used in proteomics have been available to researchers for several decades but the development of new software for analysing large data sets, and the availability of vast amounts of sequence data, enable researchers to study proteins on a global scale. The Human Proteome Organisation (HUPO) [A3.10 13] was formed in 2001 to coordinate regional and national organisations into a worldwide consortium aimed with making information from proteomics experiments widely accessible. It was feared that many of the proteome projects were being run by commercial organisations, attempting to find potential drug leads, who would patent each protein that could be identified [A3.10 15]. HUPO aims to provide public access to all the proteins that can be identified in various human tissues. HUPO also has a wider remit to assist all kinds of publicly funded proteomics efforts. The main investigations led by HUPO are the Human Plasma Proteome, Human Liver Proteome, Human Brain Proteome Project and the Mouse and Rat Proteome Project.

The challenge of discovering the proteome of all model organisms is far greater than genome sequencing, because the proteome is highly dynamic. The set of expressed proteins at any one time depends upon the conditions of the sample and the experimental methodology, unlike genome sequencing that should produce the same results regardless of the method employed. Proteins can be localised to particular organelles, cell types, organs or tissues. Furthermore, the total number of functional proteins is thought to be at least a factor of ten, perhaps up to a factor of a hundred higher than the number of genes. The “one gene - one protein” theory does not seem to hold true, as differential splicing appears to be very common, particularly in higher eukaryotes [A3.10 20], causing many different proteins to be translated from one gene. Each protein form is produced in a tightly controlled manner that is dependent on cellular signals. Many proteins are also chemically modified (post-translational modifications, PTMs), causing proteins to locate to a different part of the cell, different tertiary or quaternary structures to form, or new protein-protein interactions to form. In each of these cases, the modification is intrinsically linked to protein function.

The informatics challenge presented by proteomics is enormous because in addition to the potentially infinite size of the proteome, the experimental methods are not standardised. Different laboratories use different instruments and software, many of which have their own data format that cannot be interpreted by other applications. In contrast, genome and microarrays experiments have become standardised to some extent. Protein abundance or localisation data are dependent upon the sample conditions, and on the laboratory techniques used. Data from different techniques may not be comparable unless there is detailed reporting of the experimental protocols employed, and of the statistical processing used to normalise data points across an entire investigation.

The Proteomics Standards Initiative was formed at the HUPO congress in Washington (April 2002 [A3.10 16]) to address the data sharing challenges that arise in proteomics. The

experiments could potentially be useful to researchers working in a variety of domains but data sharing is hindered because there are no major central databases for experimental data, and no widely accepted standard formats for transferring data between research groups. The Proteomics Standards Initiative is sub-divided into three groups addressing data sharing for mass spectrometry (MS), protein-protein interactions and separation-based techniques. In the following section, the experimental techniques, and the challenges in data sharing, are briefly described. The current progress towards standard formats and the development of controlled vocabularies (ontologies) will be addressed in Section 3. The PSI has solved many of the challenges inherent in standardisation, most notably the challenge of gaining international agreement, and there are many future challenges that must still be addressed. Section 4 will expand the discussion of these issues.

2 Experimental techniques

Protein Separation

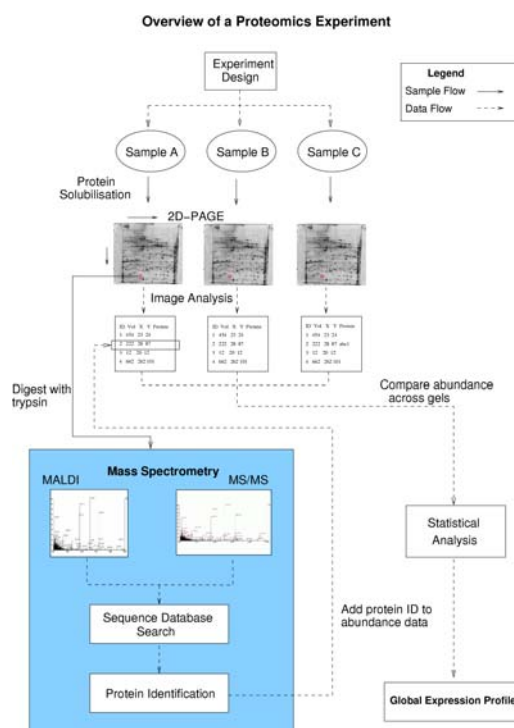


Figure 1: The data flow in a gel-based proteomics experiment.

The collections of proteins present in a sample can be detected, and in some cases quantified using a range of experimental techniques. Most techniques begin with the extraction of proteins from the starting sample. The protein mixture is then separated according to certain properties of the proteins. The most common separation techniques used in proteomics is two dimensional gel electrophoresis (2-DE), which was first pioneered in the 1970s and 1980s by Angelika Görg and colleagues [A3.10 10]. 2-DE separates proteins according to their charge

in the first dimension and their molecular weight in the second dimension. The gel is stained to allow proteins to be visualised and an image of the gel is created by scanning. The image is analysed with specialised software that detects the size of spots and estimates the volume of protein present. The software also has functionality to detect the spots on different gels that correspond to the same protein and estimate the relative difference in volume, for example to compare the proteome of a diseased sample versus a normal sample (Figure 1). The reproducibility of 2-DE has gradually improved, enabling up to thousands of discrete spots to be detected on a single gel. However, there has not been substantial research into the methods employed by different image analysis packages to estimate the relative volume of protein present on different gels, and it is therefore difficult to gain accurate quantitative measures.

There have been two novel methods reported that can quantify the relative volumes of proteins in two samples, by labelling the proteins from one sample with an isotopically heavy reagent, and proteins from a different sample with a "light" reagent, known as ICAT [A3.10 11] (Isotope Coded Affinity Tags) and SILAC [A3.10 3] (stable amino acids in cell culture). These experiments utilise the techniques of liquid chromatography for protein separation and mass spectrometry for protein identification (and quantification), which are described below.

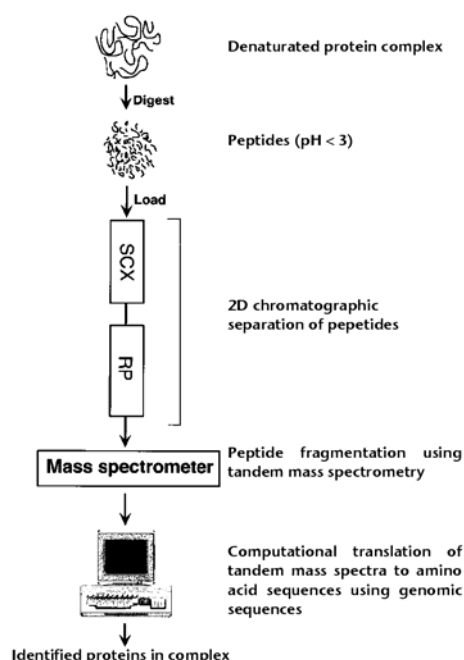


Figure 2: Two dimensional liquid chromatography coupled with MS for identifying large numbers of proteins from a mixture, reproduced from “Direct analysis of protein complexes using mass spectrometers”, A.J. Link, J. Eng, D.M. Schieltz, E. Carmack, G.J. Mize, D.R. Morris, B.M. Garvik and J.B. Yates 3rd, *Nature Biotechnology*. 17:676-682, 1999 [A3.10 17]. Two phases of LC are performed: (i) strong cation exchange (SCX) for separating by charge, (ii) reversed phase (RP) separating by hydrophobicity, followed by tandem mass spectrometry.

Liquid chromatography (LC) can be used for large-scale separation of proteins. The principle of LC is to pass the protein mixture through a column containing a liquid, which allows different proteins to pass through in a length of time that is dependent on some property of the protein, such as its hydrophobicity. Different fractions are collected over time for downstream analysis, such as protein identification. The technique has been refined to

separate at high-resolutions by utilising two sequential column stages that separate in two dimensions (2D-LC), by first separating on the different charges of proteins, then separating by hydrophobicity. The technique is powerful because a complex mixture of proteins can be rapidly separated and identified with minimal manual laboratory handling required [A3.10 33].

Mass spectrometry

The large-scale approaches have only become feasible due to the huge increase in genomic sequence data deposited in publicly accessible database, which allow proteins to be identified using mass spectrometry (MS). MS is used in proteomics in the following way. A protein spot is excised from a gel, or collected from a column fraction, and it is digested with a protease, such as trypsin. The protease cleaves the protein into peptides at predictable positions along its length. The mass of each peptide can be detected in a mass spectrometer, and the set of masses (the “peptide mass fingerprint”) can be searched against a theoretical digest of all the protein sequences deposited in databases for the closest matches. There are various software applications that perform the database searches (examples include MASCOT [A3.10 18], ProteinProspector [A3.10 27], PepMapper [A3.10 26] and SEQUEST [A3.10 7]).

Each application has various input parameters and a number of statistical measures that allow confidence to be placed on correct protein identification.

MS presents significant challenges for data sharing because there are a number of different instrument vendors, each of which is tied to a software application that produces data in a proprietary format that cannot be interpreted by any other software. Furthermore, the database search applications do not use a standard statistical model for protein identification, and some present the output of the searches only within HTML formatted web pages. Web pages are difficult to process automatically because the format is intended for expressing how the results should be graphically represented rather than expressing the types of data that exist.

Protein interactions

There are two main experimental techniques used to investigate protein-protein interactions: Yeast Two-Hybrid and affinity columns. Yeast Two-Hybrid [A3.10 8] has become a widespread technique for finding pairs of proteins that interact. The experimental basis is that the DNA binding domain of a transcription factor A is fused to protein X, and the activation domain of transcription factor A is separated and fused to protein Y. Transcription factor A switches on a gene that causes a visible change in a cell culture, causing cells to grow rapidly, or a particular colour to develop. Transcription factor A can only switch on the gene if its two domains come into contact, caused by protein X and protein Y interacting (Figure 3). The technique has been demonstrated on a large scale, identifying almost a thousand potential interactions in one study [A3.10 32]. Yeast Two-Hybrid has drawbacks that interaction partners are forced to co-localise in the nucleus of yeast cells, and therefore false positives may arise as the proteins may not come into contact in vivo [A3.10 5].

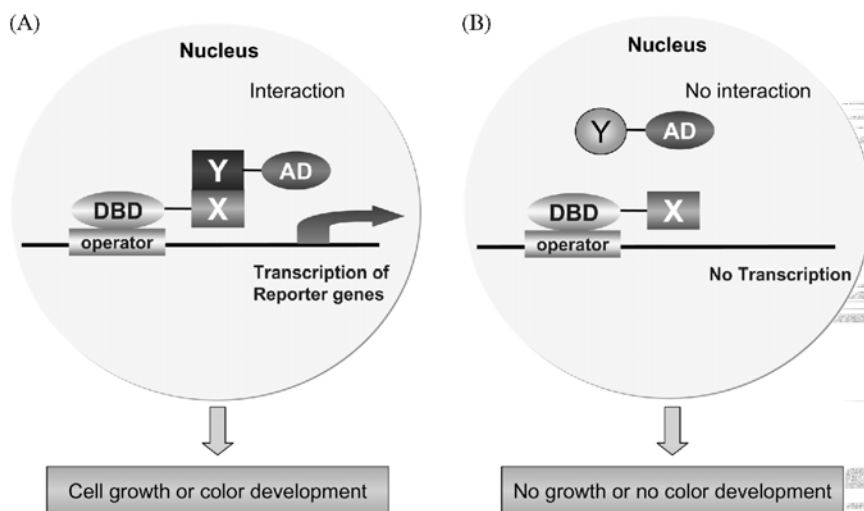


Figure 3: A summary of Yeast Two-Hybrid experiments, reproduced from “Protein-protein Interaction Networks: from Interactions to Networks”, *Journal Biochemistry and Molecular Biology*, 37:45-52, 2004 [A3.10 5]

An alternative method for detecting protein-protein interactions is affinity purification of multiprotein complexes. In this method a single protein A is fused to a tag that can be purified using an antibody. The antibody is attached to a column through which a mixture of proteins is passed. Certain proteins in the mixture interact with A, forming a multi-protein complex. The complex can be extracted, separated on a one or two dimensional gel and the proteins can be identified using MS.

Yeast Two-Hybrid and column based methods have produced large volumes of data, much of which is available in public databases. However, in the past there have been various problems integrating the information present in different systems due to heterogeneity in formats for downloading data, and no standard query mechanisms. The work of the PSI to solve some of these problems is described in the following section.

3 Data standards developed by PSI

The PSI is developing standards in three areas that are addressed in the following sections. The first is for protein-protein interaction data arising from Yeast-Two Hybrid and affinity columns. Standards are also under development for mass spectrometry, in collaboration with the major instrument manufacturers. The third area that requires standardisation has been called general proteomics standards (GPS), and covers techniques for separating proteins. There have been three official meetings of the PSI: an initial meeting at the European Bioinformatics Institute (EBI) in October 2002 [A3.10 24]; at the EBI in January 2003 [A3.10 25] and in Nice, France in April 2004.

There have also been satellite meetings at the HUPO World Congress (details available via the PSI web site [A3.10 28]).

3.1 Protein-Protein Interactions

There are various publicly accessible database that provide access to results from Yeast Two Hybrid and affinity column experiments, such as BIND [A3.10 2], DIP [A3.10 6] and MINT [A3.10 36]. There is substantial overlap in the coverage of different systems but the process of integrating data from different is hindered because of the different formats produced by each system. The PSI has attempted to alleviate this problem by the development of the molecular interaction format [A3.10 12]. The format has been developed as an XML Schema to ensure that data expressed in XML is formatted in a uniform way. The standard is being developed incrementally. The first version (level 1) covers binary and more complex interactions but does not capture substantial detail about the experimental methodology. Level 1 sufficiently covers the majority of the data that is available in the public databases, although it is a fairly simple format to enable database providers to produce output in the format as soon as possible. Future versions of the format are planned that will incorporate a structured description of the type of interaction, the interaction process, and the experimental protocol used to generate data. Controlled vocabularies will be an important component to ensure standard values are used to populate the format. The Gene Ontology (GO) [A3.10 9] will be used for describing genes and proteins in the format and the NCBI Taxonomy [A3.10 21] will be used to standardise the names of species. The work group has also formulated its own controlled vocabulary of terms. The terms are used at specific places in the format, and each term has a definition, and some terms have a link to a literature reference. The controlled vocabulary has been developed in the same format as GO, and it is available via the Open Biological Ontologies project (OBO) [A3.10 23].

3.2 Mass spectrometry

The problem of proprietary formats in mass spectrometry (MS) was highlighted above. A subgroup of the PSI has formed to model the common parts of the data formats produced by different software applications. The group also has a wider remit to describe the input parameters and the output from database search applications used to identify proteins. The first version of the format, mzData, has been developed as an XML Schema that captures the peak list on an MS trace [A3.10 29]. The format also captures the MS instrument parameters. A controlled vocabulary of terms is being developed by the PSI in collaboration with the American Society for Mass Spectrometry (ASMS) [A3.10 1]. Software is under development to convert the current output formats produced by different vendors into mzData although the majority of vendors are committed to producing output directly in mzData when the format has been officially finalised. The format will be extended to capture information about database searches carried out with MS data to identify proteins, which is an area in need of standardisation. Different software packages produce different statistical measures of the likelihood of correct protein identification, and this information is rarely published with proteomics data sets. Therefore, across large-scale analyses it is often not possible to ascertain the probability that proteins have been correctly identified.

3.3 General proteomics standards

In January 2003 there was an initial attempt to model the data that arises from proteomics experiments involving separation techniques and mass spectrometry. The model was developed at the University of Manchester and is called the Proteomics Experiment Data Repository (PEDRo) [A3.10 31]. PEDRo covers a typical laboratory workflow and it is

divided into four parts, describing: the biological sample origin, protein separation techniques, mass spectrometry protocols and mass spectrometry data analysis. PEDRo is expressed in UML and an XML Schema has been released to enable developers to create standard output in an XML file format called PEML (Proteomics Experiment Mark-up Language).

The general proteomics standards (GPS) subgroup of the PSI is creating the first version of the official standard, using PEDRo as a starting point. The standard will comprise several key components: an object model to describe the components of the standard expressed in UML (PSI-OM), an XML based markup language for data transfer (PSI-ML), an ontology for populating the format (PSI-Ont) and a document describing the information that must be made available by researchers wishing to publish, called MIAPE (Minimum Information About a Proteomics Experiment). The initial goal of the GPS group is to create a format for expressing the data that arise from 2-DE and other large-scale separation techniques, such as LC. A framework is required into which the MS format can be integrated, and ultimately the protein-protein interaction format may be described in a common format. The format must describe the experimental protocols, the biological samples used in the experiment, and the experimental hypothesis. This can be achieved by making extensive use of controlled vocabularies. The representation of biological samples may also be facilitated via collaboration between the PSI and other standardisation organisations that also require a controlled description of samples.

The MIAPE document is currently under review by various organisations and it is likely to be published in early 2005. The document will set out the requirements that must be fulfilled before a journal publication is accepted. MIAPE must therefore have significant input from the major journals. A future PSI meeting is planned where the main proteomics journals will be represented to ensure that there is good support for the requirements document and the PSI data standard.

4 Discussion

Challenges faced by PSI

The PSI has successfully overcome several major challenges in the early stages of standardisation. One of the major problems has been gaining widespread acceptance from the proteomics community. Dr Chris Taylor is head of the General Proteomics Standards subgroup of the PSI, and was a key informant in this case study. Dr Taylor highlighted the requirement for the PSI to gain support from academic researchers, industry, the major journals and from funding bodies. Many academic research groups understand the immediate benefits that will arise from being able to access the data produced from other laboratories for re-analysis, allowing new insights to be derived. Therefore, academic acceptance of the PSI has been very good, and most groups are willing to join the official effort. Industrial groups receive less benefit from standardisation because many organisations have their own "in-house" methods for sharing data, and making data publicly accessible is not a major goal. However, industry is starting to see the benefits, especially the vendors of instruments and software, as the efforts of the PSI gains momentum and their users will want equipment that supports the standard.

There have been several efforts in the past to standardise mass spectrometry due to the problems of proprietary formats described above. However, all the efforts have failed in terms of getting widespread vendor agreement because the standardisation has usually been led by one company, releasing a set of guidelines or an exchange format. Other companies are not willing to use a format produced by another group and have created rival efforts that continue to exist in parallel. The PSI MS format, mzData, looks set to succeed because it has the backing of HUPO, which is seen as an independent public organisation. The PSI has been well supported by the major instrument vendors who are all committed to creating tools that produce output in mzData.

The support from journals for PSI is essential if data sharing for proteomics is going to succeed. There are three main journals that focus solely on proteomics: Molecular and Cellular Proteomics (MCP), Proteomics, and the Journal of Proteome Research. MCP have already released a set of guidelines for authors wishing to publish MS data used for protein identification [A3.10 4]. Their guidelines are intended to ensure that other researchers reading the article can assess the likelihood that proteins have been identified correctly, using a statistical measure. MCP have been involved with PSI and there is a meeting planned for Spring 2005 at which representatives of the major journals will discuss the requirements for publication of proteomic data sets. The MIAPE document will specify the minimum information that must be made available by researchers wishing to publish a proteomics experiment. The document has been formulated by members of the PSI and will be discussed with the journals. An area that is open for discussion is the scope of the PSI in determining the quality of a study, in terms of the statistical confidence that can be placed on results. It is likely that PSI will restrict its scope to ensuring that there is accurate reporting of the methodology used, to ensure that other researchers can assess the quality of a study, and that there should be no formal requirement on the experimental procedures employed.

The major funding bodies are beginning to give support to the PSI, and it is an area that is likely to be well supported in the future. In particular, any organisation that aims to maximise the availability of data, that in many cases are extremely costly to obtain, will provide major benefits for the scientific community. It is hoped that the funding bodies recognise the requirements for computational infrastructure to solve the challenges in data sharing, which are generated by large-scale experiments.

Challenges to be addressed

It is planned that the standard under development by the GPS will act as a framework through which all three standards (molecular interaction, MS and GPS) can be integrated. This may prove to be a major challenge for the PSI due to technical difficulties, and high level problems of gaining cross community agreement. The technical difficulties arise due to the different development models that have been employed. GPS is developing a standard using an object model representation, expressed in UML, from which an XML representation can be derived. The other two standards are being developed only through an XML Schema, making extensive use of controlled vocabularies to populate the XML format. The integration of formats could occur through the development of an XML Schema that covers all the technologies but this may produce backward incompatibility problems with the versions of the format that already exist. It will also be difficult to gain an agreement on the format that can support such a wide base of potential users.

A second problem the PSI must face will be defining a structured description of the experimental hypothesis, the biological samples and the protocols. This is a major challenge because proteomics is not only a technique used by molecular biologists but is also widely used in medicine, environmental research, plant and food science. The type of biological sample on which proteomic analysis could be performed is unlimited, and it is therefore very difficult to standardise. The solution may be to create a fairly generic format that can be extended using ontologies. The ontologies will be developed by experts in a particular domain and will only be vetted by the PSI. The problem of standardising the description of samples has already been addressed by the microarray standards organisation, MGED. It is possible that PSI and MGED will collaborate to solve these challenges.

Another potential problem is allowing for the gradual evolution of the standard as new laboratory techniques are developed, which must be covered by the format. There are various solutions, such as incremental extension to a format and releasing particular versions. This approach has the drawback that developers may not wish to spend significant lengths of time developing software that can produce output in the format, knowing that it will change at regular intervals. A better solution may be to design a fairly generic format that is instantiated using ontologies that can be gradually extended, without affecting the functionality of software based on the format. In the microarray domain, developers were assured of the future support for the object model because it was controlled by the Object Management Group (OMG) [A3.10 22], who were involved in the standardisation process. OMG were able to guarantee that the first version of the standard for be stable for a specified length of time to give developers confidence that software would not be developed around a format that would become deprecated. It is possible that OMG will be involved with a similar procedure for PSI.

Integration across the “omics”

The PSI has an extended role beyond the domain of proteomics to ensure that there is agreement between the different “omics”. There is a growing momentum towards storing gene expression and proteomics data in a shared data format [A3.10 35, 14]. The biological sample from which mRNA or proteins are extracted can be described in common terms regardless of the down-stream processing that occurs. The MAGE-OM format for microarray data captures the biological samples, from which mRNA is extracted, in a fairly generic notation that could be used for other types of experiment [A3.10 30]. The PSI has begun to work with MGED (MicroArray and Gene Expression Data) society [A3.10 19] who manage microarray standards, to create a shared description of biological samples, and experimental hypothesis, that can be used by any type of large-scale biological investigation.

In addition to microarrays and proteomics, there are experiments that measure all the small molecules (metabolites) in a sample, known as metabolomics. The metabolites can be separated using techniques such as liquid or gas chromatography, and detected using NMR (Nuclear Magnetic Resonance) or MS [A3.10 34]. There is no formal organisation managing standards for metabolomics, however it may be feasible to capture all types of functional genomics data in one format, managed by MGED and PSI.

5 Conclusions

The Proteomics Standards Initiative was formed to improve the facilities for research groups to exchange and publish experimental data. Data sharing is a major problem for proteomics due to the complexity of the experimental techniques and the heterogeneous formats produced. The PSI has released the first version of a format for exchanging protein-protein interaction data between different databases. A format has also been created for representing the output from mass spectrometers that will be supported by all the major vendors, alleviating the problems of proprietary formats that have hindered sharing of MS data for many years. A standard is in development for capturing proteomic laboratory workflows, and a document will soon be released outlining the requirements that must be met by authors before a publication can be accepted. The efforts of the PSI have been well supported by academic researchers and industry. The major journals have become involved in the standardisation process, which is essential if researchers are to be forced to abide by the guidelines.

References

- [A3.10 1] American Society for Mass Spectrometry (ASMS). <http://www.asms.org/>.
- [A3.10 2] BIND at Blueprint. <http://www.blueprint.org/bind/bind.php>.
- [A3.10 3] B. Blagoev, I. Kratchmarova, S. E. Ong, M. Nielsen, L. J. Foster, and M. Mann. A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nat Biotechnol.*, 21:315–318, 2003.
- [A3.10 4] S. Carr, R. Aebersold, M. Baldwin, A. Burlingame, K. Clauser, and A. Nesvizhskii. The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol Cell Proteomics.*, 3:531–533, 2004.
- [A3.10 5] S. Cho, S. G. Park, D. H. Lee, and B. Chul. Protein-protein Interaction Networks: from Interactions to Networks. *J Biochem Mol Biol.*, 37:45–52, 2004.
- [A3.10 6] Database of Interacting Proteins (DIP). <http://dip.doe-mbi.ucla.edu/>.
- [A3.10 7] J. Eng and J. Yates. SEQUEST. <http://fields.scripps.edu/sequest/>.
- [A3.10 8] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340:245–246, 1989.
- [A3.10 9] The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res.*, 11:1425–1433, 2001.
- [A3.10 10] A. Görg, W. Postel, and S. Gunther. The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*, 9:531–546, 1988.
- [A3.10 11] S. P. Gygi, B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, and R. Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol.*, 17:994–999, 1999.
- [A3.10 12] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, et al. The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat Biotechnol.*, 22:177–183, 2004.
- [A3.10 13] HUPO - The Human Proteome Organisation. <http://www.hupo.org/>.
- [A3.10 14] A. Jones, E. Hunt, J. M. Wastling, A. Pizarro, and C. J. Stoeckert Jr. An object model and database for functional genomics. *Bioinformatics*, 20:1583–1590, 2004.

- [A3.10 15] J. Kaiser. Proteomics. public-private group maps out initiatives. *Science*, 296:827, 2002.
- [A3.10 16] J. Kaiser. PROTEOMICS: Public-Private Group Maps Out Initiatives. *Science*, 296:827, 2002.
- [A3.10 17] A. J. Link, J. Eng, D. M. Schieltz, E. Carmack, G. J. Mize, D. R. Morris, B. M. Garvik, and J. R. Yates 3rd. Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol.*, 17:676–682, 1999.
- [A3.10 18] MASCOT, published by Matrix Science. <http://www.matrixscience.com>.
- [A3.10 19] Microarray Gene Expression Data Society (MGED). <http://www.mged.org/>.
- [A3.10 20] B. Modrek, A. Resch, C. Grasso, and C. Lee. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.*, 29:2850–2859, 2001.
- [A3.10 21] The NCBI Taxonomy Homepage. <http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/>.
- [A3.10 22] The Object Management Group. <http://www.omg.org/>.
- [A3.10 23] Open Biological Ontologies (OBO). <http://obo.sourceforge.net/>.
- [A3.10 24] S. Orchard, P. Kersey, H. Hermjakob, and R. Apweiler. The HUPO Proteomics Standards Initiative meeting: towards common standards for exchanging proteomics data. *Comp Funct Genom*, 4:16–19, 2003.
- [A3.10 25] S. Orchard, P. Kersey, W. Zhu, L. Montecchi-Palazzi, H. Hermjakob, and R. Apweiler. Progress in establishing common standards for exchanging proteomics data: The second meeting of the HUPO Proteomics Standards Initiative. *Comp Funct Genom*, 4:203–206, 2003.
- [A3.10 26] PepMAPPER. <http://wolf.bms.umist.ac.uk/mapper/>.
- [A3.10 27] ProteinProspector. <http://prospector.ucsf.edu/>.
- [A3.10 28] The Proteomics Standards Initiative. <http://psidev.sourceforge.net/>.
- [A3.10 29] PSI-MS XML Data Format. <http://psidev.sourceforge.net/ms/>.
- [A3.10 30] P. T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, et al. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.*, 23, 2002. RESEARCH0046.
- [A3.10 31] C. F. Taylor, N. W. Paton, K. L. Garwood, P. D. Kirby, D. A. Stead, Z. Yin, et al. A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat. Biotechnol.*, 21:247–254, 2003.
- [A3.10 32] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403:623–627, 2000.
- [A3.10 33] M. P. Washburn, D. Wolters, and J. R. Yates III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol.*, 19:242–247, 2001.
- [A3.10 34] W. Weckwerth, V. Tolstikov, and O. Fiehn. Metabolomic characterization of transgenic potato plants using GC/TOF and LC/MS analysis reveals silent metabolic phenotypes. In *Proceedings of the 49th ASMS Conference on Mass spectrometry and Allied Topics*, pages 1–2. Chicago: Am. Soc. Mass Spectrom., 2001.
- [A3.10 35] S. Xirasagar, S. Gustafson, A. Merrick, K. B. Tomer, S. Stasiewicz, D. D. Chan, et al. CEBS Object Model for Systems Biology Data, CEBS MAGE SysBio-OM. *Bioinformatics*, 20:2004–2015, 2004.

- [A3.10 36] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni. MINT: a Molecular INTERaction database. *FEBS Lett.*, 513:135–140, 2002.

APPENDIX 4

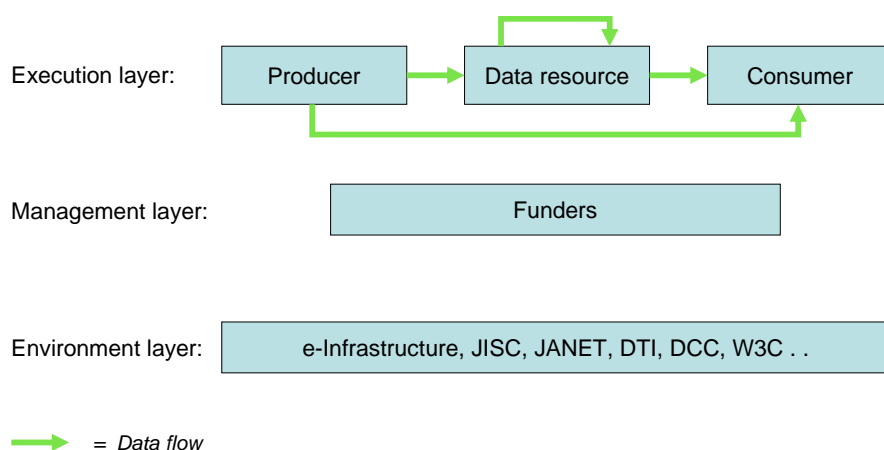
FURTHER SHARING MODELS

Models of effective data sharing can be presented from many perspectives. They include: roles, life-cycle, institutional, information flow and funding models which support sharing. No single perspective can present all aspects of sharing adequately. Funding and data flow models are discussed in section 15 of the main report. We present here models from the first three of these perspectives.

Objectives, needs and responsibilities

In this section we examine the roles of the major players in the data sharing process and tabulate their respective objectives, needs and responsibilities. The objective of presenting this model is to assist identification of where action may be taken, and by whom, to improve sharing processes.

We distinguish three major layers in the process:



The execution layer – objectives¹⁶, needs, functions of producers, data resources and consumers

The following table (overleaf) presents producers, data resources and consumers from left to right as in the top level of diagram above following information flow. In terms of needs for effective data sharing, however, requirements tend to flow from right to left, from consumer to producer; in commercial terms “giving the customer what he or she wants”. The second row of the table summarises data-sharing **objectives** for the three actors in the sharing chain.

Producers	Data resources	Consumers
-----------	----------------	-----------

¹⁶ For consumers the actual objectives, for producers and data resources the desired objectives

Objectives	To receive: Recognition and reward for contributions to sharing	To be recognised for: Service delivery to their designated communities An ethos of high quality service (whilst fulfilling any other, scientific, goals)	Have access to data and tools that are: Relevant Trustworthy, reliable Easy to find Easy to incorporate into their own systems and to use Fully and accurately described (metadata) Free, or at an affordable price
-------------------	--	--	---

The next table summarises the **responsibilities** of each of the three actors to achieve effective data sharing, and thus to fulfil the objectives summarised above.

	Producers	Data resources	Consumers
Responsibilities	Undertake data planning, starting at grant application stage Adhere to approved data plans Produce and pass on data in sharable standards Provide metadata early, of good quality and in sufficient quantity	Make data/metadata available Provide tools Data and tools quality assurance Manage data and metadata; undertake curation, provide added value Provide assistance & training Administer collections overall Continue consumers' data plans & undertake further planning	Re-use existing data rather than re-create new, unless there is a clear scientific need Respect conditions on data use: IP rights, security, confidentiality, and ethical constraints

The third table below summarises their **needs** to enable them to meet the responsibilities shown above.

Producers	Data resources	Consumers
------------------	-----------------------	------------------

Needs	<p>Resources to prepare data for sharing, including planning, and metadata production</p> <p>Incentives to undertake the above responsibilities</p> <p>Assistance, training, support for these tasks</p> <p>Awareness of consumers' needs</p>	<p>Clear strategic objectives</p> <p>Adequate long-term funding</p> <p>Career structures which encourage staff retention and reward</p> <p>Training for curators and data managers</p>	<p>Know data may exist</p> <p>Know how to find data (discovery tools)</p> <p>Data in usable standards</p> <p>Metadata to assist interpretation, use, processing</p> <p>Quality and provenance indicators</p> <p>Rights information</p> <p>Assistance, support</p>
--------------	---	--	---

The management layer – policies and actions

The middle, management layer in the diagram data flow diagram above represents the funders. Their objectives can be summarised by:

Making optimal use of expensively gathered data

Encouraging first-class, productive science.

In overall terms the requirements are to recognise and encourage the objectives of the three groups in the sharing chain, provide for the needs which are articulated, and monitor that responsibilities are shouldered, (as set out in the tables above). Specific requirements and areas to influence within the three sharing groups are as shown in the table on the following page.

The environment layer – strategic and international engagement

As noted above, the funders and the sharing chain operate in a complex environment of interacting factors, and many of these are beyond the study sponsors' influence. A few general points can be made:

This study is evidence of funders being willing to coordinate their approach to sharing. As pointed out, sharing transcends barriers – institutional, national and between disciplines. Clearly continuation of coordination is required to make an impact in an environment of interacting players. Coordination is needed also with like bodies abroad.

Science and technology change rapidly, calling for monitoring of developments which may either hinder or promote contemporary sharing practices.

Guidelines are needed for handling those sharing data in the context of collaboratively funded and/or collaboratively executed projects.

Mechanisms to resolve differences between different funders where they are at variance over collaboratively produced data.

Table A1: funders' responsibilities to producers, data resources and consumers.

To producers	To data resources	To consumers
<ul style="list-style-type: none"> • Encourage the development of reward structures for sharing: E.g.: - Data citation - Promotion board briefings • Insist on data planning for sharing as a condition of grant • Provide resources (including time) for the proper preparation of data for subsequent use by others • Support awareness raising initiatives regarding the scientific and career value of data. E.g.: - Publicise success stories - Bulletins, staff newsletters • Provide support facilities. E.g.: - Help desks - Training courses - Materials on intranets • Monitor performance 	<ul style="list-style-type: none"> • Articulate strategy and policies with respect to data sharing • Establish explicit service delivery goals for data curation and tools development • Set service-level agreements and norms; monitor performance against these • Separate more clearly service delivery functions from research • Introduce longer-term funding horizons for data resources • Establish career paths for those engaged in the management and curation of data and tools, including training opportunities • Brief promotion boards on the value of service careers for curators and data managers 	<ul style="list-style-type: none"> • Unless there is scientific need (such as validation) do not fund research which repeats data collected elsewhere • Provide guidance materials on the legal, regulatory and ethical use of data • Provide support facilities E.g.: - Help desks - Training courses - Materials on intranets • Support awareness raising of the value of data re-use among the consumer group(s) to which one has access. • Engage with those doing research into curation and preservation questions.

A life-cycle approach – data management planning

In Chapter 8 we recommended a data management plan, to be developed by data producers at the project planning phase, before the award of funds. In this model we elaborate on that proposal, following data from creation to destruction.

Why a plan? There are two fundamental reasons:

14. Data is rarely “self-standing” – it requires further information to be provided in order that it can be used effectively, particularly in the later stages of its life-cycle when tacit knowledge about it has been lost or forgotten. The need for further information is always the case when data is in digital form, for otherwise it can only be viewed as a string of bits with no further meaning: at the very least you need to know what program to read it with and the type of machine on which the program runs. This implies that metadata has to assigned to data, processes may need to be recorded, infrastructures for its storage and use specified. These require planning, not just at project plan level, but also at a higher level, with metadata fields agreed, co-ordinated, for inter-operability of metadata.

15. Efficiency, both during the creation and initial exploitation phase of data use, as well as for its future storage, curation, and re-use in a sharing context.

Our recommendation for data planning recognises that funding institutions and host research organisations will have (or should have) objectives for the data they fund, and strategies and policies for its exploitation. Therefore we believe that specific data plans should be guided by these, and that they should be assessed at or before the point of decisions about funding, and that they should originate from the potential data producer.

What should a plan contain? Its scope should cover the expected complete span of the data life-cycle. Suggested topics for a plan are:

Technical standards to employ, processes to be applied, analyses to be performed, storage requirements, volume estimates, software and other tools to be used for collection, analysis, visualisation and reporting (including programming languages, if programmes are to be written)

What metadata needs to be collected, how, by whom and when – not only for the collection phase but for subsequent phases through to archiving. It includes documentation of the data, of any programmes written and of processes employed in collection and analysis; data models/schema.

Resources required – including staffing and funding requirements

Security requirements, confidentiality requirements, access controls

Rights information

Potential future relevance/use of the data; definition of the “community of interest” when shared; stakeholders in the data; possible economic benefits

Estimates of any special resources needed to preserve the data and metadata.

Project management planning as regards data handling.

Generation of plans should be approached positively (AHDS is a good example of this), and producers and executors of plans will need support, advice and possibly training (again AHDS provides an exemplar; so does the EnvGen, now the NERC Environmental Bioinformatics Centre). Before advice can be taken people need to know it is available and how/where to get it, implying a need for both funding bodies and to advertise and promote these services.

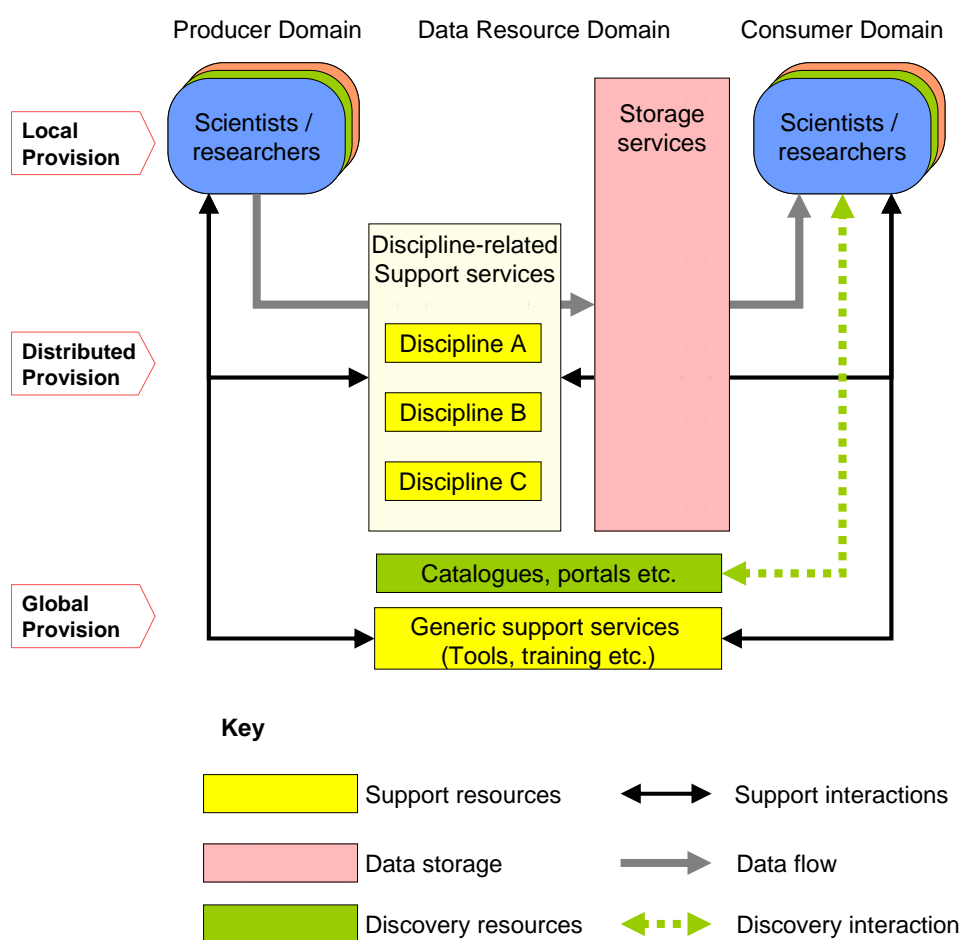
Plans should travel with the data to repositories. It is suggested that plans should be reviewed by the potential repository when they are developed so that it can ensure curation and preservation needs are adequately covered; alternatively the plans should be supplied to the repository in advance so that it can undertake more effective capacity planning.

For long projects, plans are likely to need to be reviewed, adjusted, updated.

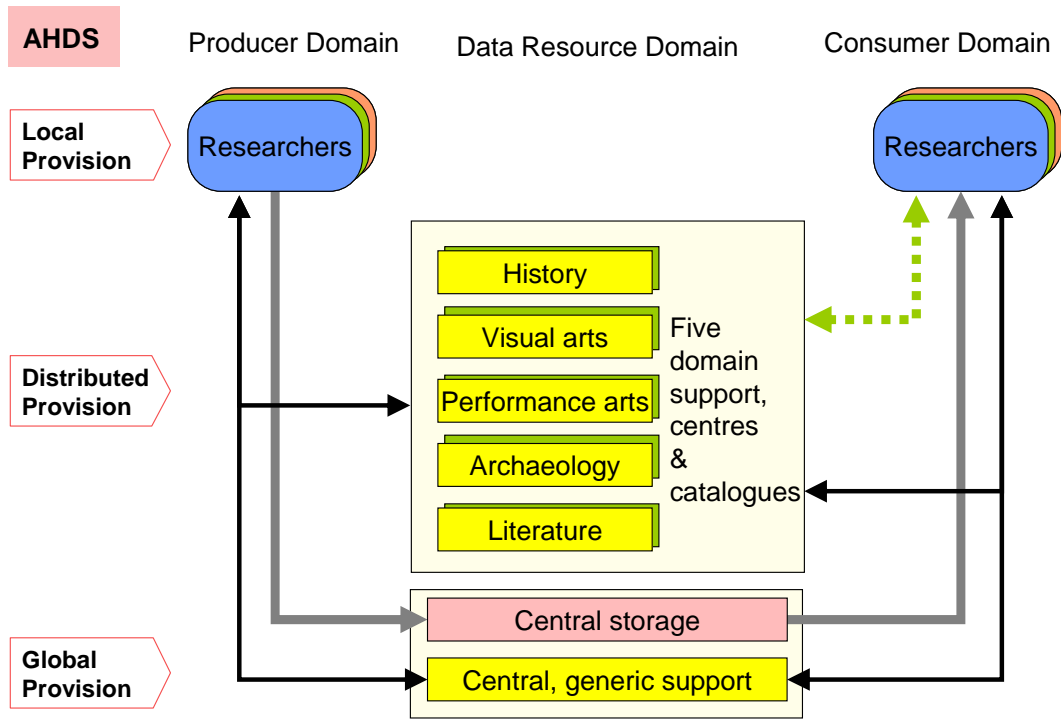
Setting criteria for selection of data for longer term retention is still an evolving art; but plans such as these may prove useful in assessing data for long retention.

An institutional approach – what happens where

The various stages and centres of responsibility required in the data sharing process were discussed in the e-Science Curation report [69]. The following general model has been adapted from that report, placing the basic functions of data storage, support and providing discovery resources in a matrix of actors (producers, data resources and consumers) and location (local to the user, distributed or centralised). The following diagram shows the generic model: storage can be local, distributed or centralised so long as it is well managed. (We have reservations about the long-term consequences of the local solution). Support has two components – discipline specific – which is probably best done in distributed specialist centres, and generic which is probably most cost effectively done centrally. Access methods – portals etc. - for discovery can be either distributed or central.



The following diagram shows how this general model is made concrete in the particular case of the AHDS, where discovery services are distributed and associated with support services for the five subject domains, while generic support and storage is centralised. Similar diagrams can be drawn for the other data resources studied.



APPENDIX 5

SUPPLEMENTARY INFORMATION

5.1 The data provenance issue

Why is provenance tracking important?

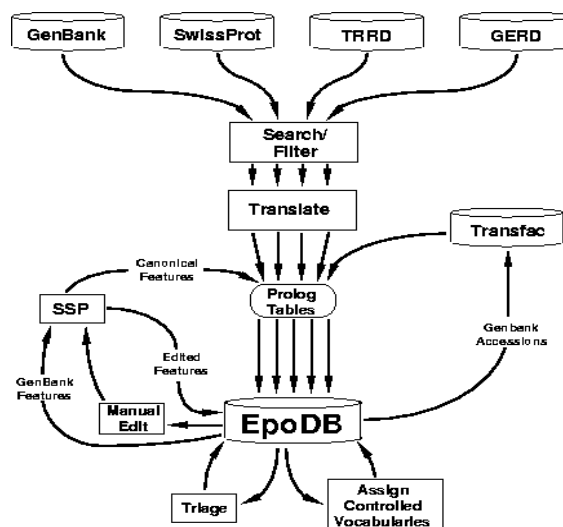
Provenance represents essential metadata about where data originates and how it has been modified. A data file is of little use if we do not know how it is formatted, similarly, the same file is equally useless if we don't have any information about where it has come from: does a file contain validated data from a scientific instrument or test output from an early (perhaps buggy) version of a simulator?

Provenance data can be thought of in two categories: derivation data and annotations. Derivation data provides the answer to questions about what initial data was used for a result, and how was the transformation from initial, raw data to result achieved. A complete derivation trail would guarantee that we can reproduce the workflow from raw data or simulation to final results. In principle this information can be obtained, stored and organised automatically. Derivation data can include information about who was responsible for entering a dataset into a repository or configuring and executing workflows to process data further. Annotations are supplementary comments supplied by the human scientists which elaborate on the purpose, circumstances or possible implications of a workflow or a results set. Annotations are usually attached to one or more such objects and can include semantic links to other annotations, results sets, work flows etc.

Why is provenance tracking hard?

In recent years we are seeing a huge growth in both the complexity of the data flows that scientists have to use and the size of scientific collaborations. Tracking data provenance is therefore becoming more important and more challenging. Fortunately these large collaborations require software support, and this supporting software can help.

In molecular biology, where data is repeatedly copied, corrected, and transformed as it passes through numerous genomic databases, understanding where data has come from and how it arrived in the user's database is crucial to the trust a scientist will put in that data, yet this information is seldom captured completely. For example, Illustration 1, taken from [3] shows how various databases are involved in the construction and maintenance of EpoDB, a database of genes that relate to vertebrate red blood cells.



The data flow of the EpoDB(Erythropoiesis Database)

Notice the complexity of this diagram. All the databases in this diagram are curated and many of the arrows involve human intervention to correct errors in the database. Data in the EpoDB is derived and translated from multiple sources. Also notice that the diagram contains more than one loop. Despite the fact that these databases are curated and great care is taken to document all the changes and corrections made to the data, establishing the complete provenance of data obtained from the EpoDB is far from trivial, and the curators themselves are concerned about this provenance problem.

In astronomy, useful results may have been obtained by filtering, transforming, and analyzing some base data by a complex assemblage of programs, yet we lack good tools for recording how these programs were connected and the context in which they were run.

The potential volume and complexity of provenance information should not be underestimated. Take, for example, the provenance needs of the high energy physics community. The volumes of data produced in experiments like the CERN Large Hadron Collider demand international collaborations on an unprecedented scale just to analyse the data. The data on which science is actually performed are derived from the experimental results through a complex series of filters and translations which include large volumes of simulated data. In order to ensure validity of the final scientific result it is important that the workflow that led to the derived data, which may have occurred over several years, can be reproduced exactly. To guarantee such reproducibility over such a time course we need to track not just the applications used and their configurations, but also the environment, operating system and hardware in which they were executed. Furthermore, we need to be very specific about the version of each piece of software that was used, the versioning information includes not just the source code that was used, but also the version information of the entire compilation environment in which it was compiled and all the libraries that it links to. So the complete provenance trail of the data would include a complete provenance trail of all the software that manipulated that data. It is quite conceivable that a complete provenance record for a complex data flow could significantly exceed the size of the data itself.

Provenance tracking raises issues in related areas such as archiving. Some important open archiving issues that have yet to be resolved include:

citation mechanisms – how can we provide robust mechanisms for referring to portions of a dataset. It has been argued that Xpath is too complex a language which does not take advantage of the structure of a well organised database.

citation preservation – even if we had the ability to produce well defined links preserving those links through time across changing data sets presents further open challenges.

What else can provenance tracking do for us?

The importance of provenance goes well beyond verification. Provenance data can provide management information about what has been done in a virtual organisation: who last used a particular workflow, and in what context, what have others done with similar data. Provenance provides a body of experience that can be used to guide further action.

Provenance may even be used in data discovery. Knowing the provenance of a data item may help the biologist to make connections with other useful data. The astronomer may want to understand a derivation in order to repeat it with modified parameters, and being able to describe a derivation may help a researcher to discover whether a particular kind of analysis has already been performed.

The annotation aspects of provenance are important value added metadata. Researchers do more than produce and consume data: they comment on it and refer to it, and to the results of queries upon it. Annotation is therefore an important aspect of scientific communication. One researcher may want to highlight a point in data space for another to investigate further. They may wish to annotate the result of a query such that similar queries show the annotation.

The MyGrid team identified a number of additional benefits to having accurate available provenance data:

It provides a store of know-how, a means of learning from history and disseminating best practice.

There is substantial benefit in using provenance to give scientists a view across multiple experiments: “Who in my community has worked on similar data, in the last six months, and can I adapt their workflow descriptions to my current experimental context?” Such facilities would give e-Scientists the ability to share community knowledge, best practice and know-how.

The web of experiment data-holdings enabled by provenance information allows a variety of personalised views to emerge: This kind of information allows views to be taken from experimental, user, data, organisational, project, etc. stances.

Scientists always wish to know if the experiment they wish to run or the hypothesis they wish to test has been performed or explored before - *in silico* work should be no different.

Provenance information can be useful from a management point of view: “What are the most used bio-services in my organisation, group or project? Is it worth renewing my subscription to this expensive service?”

In the volatile world of bioinformatics such information could be used to automatically re-run an *in silico* experiment in the light of a notification of change in data, analysis tool or third-party data repository.

What tools exist to help?

This sections looks at three examples of scientific projects, MyGrid , GriPhyN and SAM, that provide supporting software to assist in the problem of provenance tracking.

MyGrid

“MyGrid aims to deliver a personalized collaborative problem-solving platform for e-Scientists working in a distributed environment, such that they can construct long-lived *in silico* experiments, find and adapt others, publish their own view in shared repositories, and be better informed as to the provenance of the tools and data directly relevant to them. The focus is on data-intensive post-genomic functional analysis.”

MyGrid is currently developing a lab notebook application intended to allow the capture of provenance as free form text from the scientists augmented with active links to objects stored in the MyGrid Information Repository, including things like data and workflows.

GriPhyN

GriPhyN is the Grid Physics Network its mission is to “Enhance scientific productivity through discovery and processing of datasets, using the grid as a scientific workstation”.

The GriPhyN project is very interesting in this context because their technology based around the concept of “Virtual Data” has provenance tracking built in from the outset. Data access and processing within GriPhyN are described using the Virtual Data Language (VDL) which describes work flows in terms of data dependencies. A complete work flow represents a detailed recipe for the generation and/or processing required to produce the resulting data. These recipes are then stored with the derived data along with details of when and what was actually executed, to provide a detailed provenance track.

The system allows for caching and replication of data and results so that storage, network and computational resources can be traded off against each other. In most cases the user will not care how their results are produced so long as they meet the detailed specification they supplied. However, when it comes to validation and reproducibility it is important that the provenance records exactly how the data was produced.

SAM

The Scientific Annotation Middleware (SAM) [8] system is an infrastructure that includes a data storage system that provides a powerful annotation system for human comments and automatic metadata tracking, that it stores information like creation time, source and workflow about the stored data. SAM includes tools for searching, manipulating and presenting this metadata in ways that are meaningful to their users.

The system presents researchers, applications, problem-solving environments, and software agents with a layered set of components and services that provide successively more

specialized capabilities for the creation and management of metadata, the definition of semantic relationships between data objects, and the development of electronic research records. Researchers access the system through a notebook interface, available via desktop computers and mobile devices, as well as through SAM components embedded in other software systems. SAM supports manual and programmatic queries across entries generated by these multiple sources.

5.2 Standards and data sharing

This appendix is an expanded version of section 6 of the main report.

Key informants stressed the importance of standards for data sharing: “standards are essential for data sharing”, “standards make it easier to do research”, “standards and standards bodies are crucial to data sharing”. Why is this? What standards are involved in data sharing? What do they entail? How are they developed, disseminated, adopted, maintained? What obstacles and problems are confronted?

Reasons why standards are needed are shown in the following two quotations (our bolding):

*“The need for the development of standards for clinical trial laboratory data is painfully apparent to those involved in the multiple laboratory data transfers that occur during a clinical trial. **Some central clinical laboratories currently support up to 1200 different data formats. Every sponsor has developed a nearly proprietary data standard, and often has multiple standards across different therapeutic areas or business units. Each data standard requires development, validation, testing, support, training, and maintenance.** For each player in the laboratory chain ... effort is expended in exporting data from one system, importing it into another system, verifying the transfer, correcting problems, retransfer of data, and appropriate quality control.” [CDISC]*

The mass of formats multiplies work and is a massive obstacle to sharing of data with others, in particular outside an organization. Chris Angus, writing when at BT, noted that lack of standards had an impact on the quantity (of information) available:

*“Effective management of engineering information is hampered by the different computer systems, application programs and data formats used to create and store it. **These incompatibilities affect quantity [...], quality [...], exchange and sharing. [...]** There are several ways of overcoming these problems. **The ideal way is to use a standard way** of representing engineering information in a form that is computer sensible (so that it can be exchanged and shared electronically), neutral (independent of commercial hardware or software), and international (rather than company, vendor, .. national). With this “common language” you can exchange and share high-quality information between organizations and across the life cycle.” (Chris Angus, BT)*

CDISC is aiming to achieve multiple benefits from the introduction of clinical trial data standards – increased efficiency, savings in time and cost, and also streamlining the approval time for new products by the drug regulatory agencies.

Developing standards, however, is a difficult and slow business. An examination of standards encountered in this study shows that developing standards for technical data in particular is difficult, with common features of successful and unsuccessful standards.

STEP: from detail to data model

The context in which Chris Angus was that of STEP (ISO 10303), the international standard developed for the exchange of technical product data. It is estimated that STEP accounts for some two thirds of ISO activity. The evolution of the STEP standard is instructive.

During the 1970s several national protocols and standards were developed to address the then burning issue of exchanging geometric data between different CAD (computer-aided-design) systems. By the early 1980s it was clear that a single more comprehensive standard was needed, one which was international in scope and ideally should cover the whole life-cycle of use, and archive storage, of product and project data.

So in 1984 the International Standards Organisation (ISO) established TC 184/SC4 (Technical Committee 184, Sub-Committee 4) to undertake the work. Initially the committee consisted of three people, growing over the next ten years to some 500 persons from about 20 nations. It took ten years before the first 12 parts of the new standard ISO 10303 had passed through the various committee processes and were ready for publication as full international standards (December 1994). The standards produced were complex and detailed.

In 1992 the DTI funded a three-year project, with matching industry funding from some 30 companies, the project continuing thereafter on a voluntary basis. From this work, the participants realized that they needed a generic data model. This has led to ISO 15926, Industrial automation systems and integration -- Integration of life-cycle data for process plants including oil and gas production facilities. This is a seven-part standard, incorporating a Data Model and Reference Data Library (RDL) - "a language for engineers that are processable by computers" - and this is being adopted back into the STEP community.

"Once groups realize that all you have to do [to comply] is to produce your own RDL, they realize, "this is easier than they thought", and they apply it". (Key informant)

Features for success –ease of use and flexibility

So a key feature of the ISO 15926 standard is ease of use. Another is flexibility – which may seem a contradiction in terms. However, standards for data, particularly those involved in processes, must bridge a wide range of different users, uses, and technologies, so flexibility is essential.

Lack of flexibility was identified by the CDISC LAB study [#128] as one of the reasons why existing standards were not used in clinical trials:

"The structure of existing standards often does not accommodate the specific requirements of clinical trial laboratory data; [...] Some of the existing standards have limited flexibility that makes adaptation for clinical trial use cumbersome".

... leading to a proliferation of newly created digital islands - each company developed its own standards for laboratory data generated in clinical trials.

A closer look at the characteristics targeted by the model being designed by the CDISC LAB group for a laboratory data standard for clinical trials shows that more than one type of

flexibility is being targeted – flexibility for a breadth of users, and flexibility to keep pace with evolutions in laboratory testing in terms of model and technologies.

Those standards encountered during the study with detailed specifications, rather than high-level generic information models, met higher levels of resistance and poor adoption. As with the early versions of STEP, the mmCif standard developed was extremely powerful, but because it was complicated, it required effort and expertise.

As we have seen, however, re-using data generated by experiments means having sufficient information to do so, which can mean knowing about a very large number of variables (see for example the Proteomics Standards Initiative case study).

Standards – stringency and pragmatism in development

Another cause of failure was highlighted by more than one key informant, pointing to the risk that trying to produce a standard which incorporates everybody’s wishes actually produces something which is either over-complex or:

“If you try to produce something which does what everybody wants, you find that when you want to use it, it doesn’t do what you want it to do.”

Slack control of the definition process can end up with bloated standards:

“[...] you end up giving everybody the right to add, but not to veto. [...] you need to gather requirements and prune ruthlessly”.

“[...] (they did) a very great deal of work, there were lots of key players, so ended up with something that is big”.

We found that adoption of standards was higher when the standard was not too complex.

So the standards development process needs to apply pragmatism and some rigour in standards definition. In a young community, where representative organs are young and still establishing authority, this may not be easy.

Categories of standards

Type	Characteristic	Prescriptive?	Example
Informal	Used by a wide community but not formalised in a recognized institution of any sort	No	Microsoft Word, Affymetrix technology
Formal	Available to a wide community and subject to a process of definition and maintenance within some recognized institution.	No	XML
Professional	Apply to a specific, defined community and are defined and maintained by a professional body or similar.	Yes	General Medical Council
Legal	Apply in principle to a wide audience within a legislative boundary and set by a body with legal power of enforcement.	Yes	Drive on the left in the UK

Informal standards tend to be acknowledged as such without ever having gone through the wider consultation and certification processes. Microsoft Word or the Affymetrix array technology are examples of such standards, which enjoy their status because these products are very widely used – and therefore documents and data presented in these formats are easily exchanged, but only while those formats are current or recent.

However, as one key informant pointed out, industry standards also tend to be “semi-autocratic”: the manufacturers can simply make a change, and users have to accept or go without. On the other hand, companies whose products hold such commanding positions in markets tend to make the competition authorities uneasy, because monopoly positions make it more difficult for competitors to sell their products, so customers have less price power and access to quality and innovation may be stifled by the dominant supplier.

They are *de facto* standards – arrived at by a combination of practicality and market forces.

They can be contrasted with **formal** standards, whose definition and maintenance are managed by a recognized institution. A good example of the latter is W3C, the world-wide web consortium, whose RDF and OIL standards were discussed above. XML is an example of a formal standard (indeed, it has mushroomed into multiple sub-standards, perhaps so many that the sheer number of variations makes the standard more difficult to use and to manage).

In the context of data sharing **professional** standards are more likely to be of relevance in considering those who set the standards, rather than the standards themselves. They are typically defined and maintained by a professional body for a specific community, and will frequently include a code of conduct. However, it may be the case that, as with the accountancy community, one industry may be governed by one standards-setting authority which has the consent of a number of different professional bodies.

Legal standards are significant in that they carry greater sanctions than other standards, and those sanctions are enforced by the state rather than the standards-setting bodies.

5.3 Databases

Databases figure prominently in many of the case studies, for example:

The Malaria *Plasmodium falciparum* case study looks i.a. at database development designed for wide, user-friendly use, the use of the Genomics Unified Schema database platform, and an example of a curated database

BRIDGES is both user and provider: it documents its experience as a user of community databases, and its use of tools such as OGSA-DAI (Open Grid Services Architecture Database Access and Integration, which is working to overcome database integration problems), OGSA-DAIT and IBM's Information Integrator tool

GIMS (Genome Information Management System) again is both user and provider, as user of data from community databases, and designer and provider of a data warehouse (object-oriented database) supporting powerful queries.

Here we highlight some of the findings from these case studies and interviews with key informants.

Different types of databases

Databases are collections of data, organized so that it can be accessed, managed and updated with ease. As such, they are one of the most prevalent tools used for both housing and organizing data held in community repositories. The many databases housing resources, whether community or otherwise, vary in format and access methods.

Databases can be classified according to content type or according to their organizational approach. The different database types have different strengths and capabilities. Further, there are many versions of the various types available to users – commercial (more robust, but cost money), open source (cost much less, maybe have provider support, perhaps cost nothing.), free ware.

The most common type of database nowadays is probably the relational database (RDMS), a tabular database in which data is defined so that it can be organized, reorganized, accessed in a number of different ways. Some databases still use a flat-file system (where the records in the database have no structured inter-relationship), as they support searching, analysing, comparing of nucleotide or amino acid sequences.

Object-oriented databases are less common, but have advantages where storage of objects is concerned. In an object-oriented database, data is defined in object classes and sub-classes. The GIMS database is object-oriented in form, allowing it provide the user with much more powerful querying and analysis functionality. In the genomic domain, AceDB is both object-oriented and relational; written for the *Caenorhabditis elegans* genome (a soil nematode about 1mm long), it has been used for almost all plant and animal genome databases, and is good for encoding laboratory knowledge into a data structure, though now tending to be superseded.

A recent arrival is the XML database. XML is a flexible way to create common information formats and share format and data between networks, on the Web, and elsewhere, and which has been formally recommended by the World-Wide Web Consortium (W3C). The XML database format helps with database integration, by applying query languages such as Xpath and Xquery, XSLT (extensible Stylesheet Language Transformations) and SOAP (Simple Object Access Protocol), to cope with the heterogeneity of different databases. However, there are disadvantages to XML. XML files can be very large. One experience quoted by a case-study interviewee was a download of an XML database which took three and a half days.

Whichever type, database methods are perpetually being enhanced, innovations introduced, and tools developed. For instance, we now have databases which are based on ontologies.

Web access to databases – advantages and limitations

Most public databases can be searched via the web. One factor behind this is demand, either on the part of users or on the part of journals; for instance, web access is a condition of inclusion in the Nucleic Acids Research database list.

Access interfaces (usually using database connectivity tools rather than HTML) allow users to query databases, using query languages such as SQL (structured query language, a standard language for making interactive queries on the database, and updating the database), or Object Query Language. Interfaces are also used for granting access to search methods such as BLAST or FASTA.

The Sloan Digital Sky Survey (SDSS – a project mapping in detail one quarter of the entire sky, determining the positions and absolute brightnesses of more than 100 million celestial objects) is one of the few archives in its domain which allows SQL queries. Most others restrict users to a web-form interface or simple transfer of ASCII or FITS (Flexible Image Transport System, a common astronomy file format) files. As an interviewee said,

“the difficulty with a forms-based interface is that it is hard .. impossible! to ask questions that the interface designers did not think about. So while forms can provide a useful interface for the bulk of the interaction, the really cutting-edge things will require more flexibility than that.”

However, where data is held behind firewalls, these types of querying and searching can be impossible because access is only allowed via HTTP and FTP ports. This problem can be solved either by an institution providing databases on the other side of its firewalls, as in the case of Ensembl, for instance, or by using SOAP together with WSDL (Web Service Definition Language), which uses HTTP. Several of the major community resources have been adding SOAP access to their databases. However, interviewees confirmed that many researchers fear damage to their databases if they place them outside firewalls. One interviewee suggested that they place “toy” versions outside the firewall, keeping their original safe inside a firewall.

A friendly interface is key for users, and familiarity with the Web encourages use by scientists with less computing experience. There will be times, however, when users will need to interact in more complex ways with resources – for example in 3D. Several areas, and several tools can be highly complex and non-intuitive, so richer interfaces are needed. Achieving

rich, easy-to-use Web-based interfaces is very difficult. Those available on the Ensembl site are the result of much work by extremely skilled computer scientists and mathematicians.

Such interfaces make life very much easier for users. Indeed, in the BRIDGES case study, some reluctance was seen on the part of users confronted with an interface which did not resemble the Ensembl interface.

The web format for the Ensembl interface means that users (particularly those less familiar with databases) tend to phrase their enquiries in terms of the web presentation.

Another repercussion of the apparent simplicity of interfaces is the risk that funding agencies might think they are easy to build, and therefore do not require much resource, support and therefore funding. That is not the case.

Databases are difficult to integrate and change

Specialist bioinformatics units write their own wrappers and scripts to bring heterogeneous databases together, re-using these when they update their local copies of databases. However, as our interviews and case studies show, even expert database practitioners encounter problems when integrating databases or data downloaded from community databases.

Some difficulties arise simply because of differences in computing platform, environment, machines between provider and user. Difficulties such as these impeded exchange of databases between institutions during the UK's Foot and Mouth crisis in 2002, taking several days to resolve.

Projects such as the current set of OGSA-DAI initiatives may provide a substantial aid to users here.

Another difficulty lies in the fact that databases change over time. Not only do they occasionally change in form, but the organizational schema may need frequent adjustment to accommodate new scientific knowledge, performance improvements. Query interfaces can change. For the user, the terrain is ever changing, and the demands on his/her time are substantial.

The task is made particularly difficult and frustrating when schema information is not given or schema changes are not flagged. This problem was mentioned by several interviewees, for instance:

“The group wanted to obtain QTL (quantitative trait loci) data from Jackson Laboratories. They had had difficulty guessing the database schema, which appeared to be undocumented. They requested schema documentation from Jackson and were given incorrect information. Later a Google search located documentation by a database developer at Jackson Labs. The documentation described the schema. Using this information the team was able to import the data. Clearly the data providers did not even know they had the documentation required by their users.”

Because users often do not know when schemas change or databases are updated, they have to spend time checking to see whether there has been any change. Of course, this is not a technical but an administrative problem, though it can be facilitated by tools: Some resources

provide a “push” service, and more tools are emerging which would facilitate notification to users of change in schema.

Many databases provide hypertext links between entries in their databases and those in other databases. These links, which usually use database accession numbers (created for a new entry), are a common means of integrating database.

Databases, volume and performance

Volume or capacity can be a problem in various ways. Your database system may be compatible with the public database, but your computing platform may have different “heaps”, which may crash downloads. High-resolution image files tend to be large; when aggregated, this may strain performance.

For resources such as Ensembl, which is a connected set of databases continually adding more, for the computing system to cope with the multiplication of links each additional genome database calls for considerable computer systems and database design skill, or state-of-the-art computing power. (The Ensembl team stress the key contribution made by the Sanger Institute and EBI computer systems team in supporting their resource.)

Farms of unconnected databases, or databases which are just federated, however, present obstacles with regard to interoperability (including interoperability of tools to make use of the data) and cross-searching when the data are diverse. BADC has put together a set of tools for users, in collaboration with users, in this regard.

Curated databases and the propagation of error

More than one interviewee referred to the risk of propagation of error through database annotations.

“Almost all assertions about genes are made by comparison with a database [...] you then put your assertion into a database [...] blunders are then propagated through other databases ..”

Curated databases such as the Sanger Institute *Plasmodium falciparum* database, which checks entries for accuracy and keeps links up to date, are at much lower risk of this problem.

The question of tracking records which are copied from one database to another is addressed in Appendix 5.1 on provenance.

Designing and managing databases

The malaria *Plasmodium falciparum* case study looks at database design and management in more detail. Points to highlight, and reiterated by several key informants, is that database design takes a considerable time – estimates coincided at two to three years. Hasty database design results in an architecture which will call for reworking, but in the meantime the database has been populated, so the net result is very much more work.

The wheel does not need total duplication. As with many resources reviewed, case studies could turn to previous tools and resources for components and guidance. The Sanger Institute *Plasmodium falciparum* database uses the Genomics Unified Schema, a collaborative project between six institutions.

The Ensembl resource tries to plan databases over three years; it works on projects first as pilot projects, and is prepared “to throw away code”. It identifies three aspects to database management – content, function, and management, with the two former driving the latter. Running the database is far more complicated than just a matter of the way the data works. Achieving a balance between how much the resource develops and maintaining the resource is difficult.

