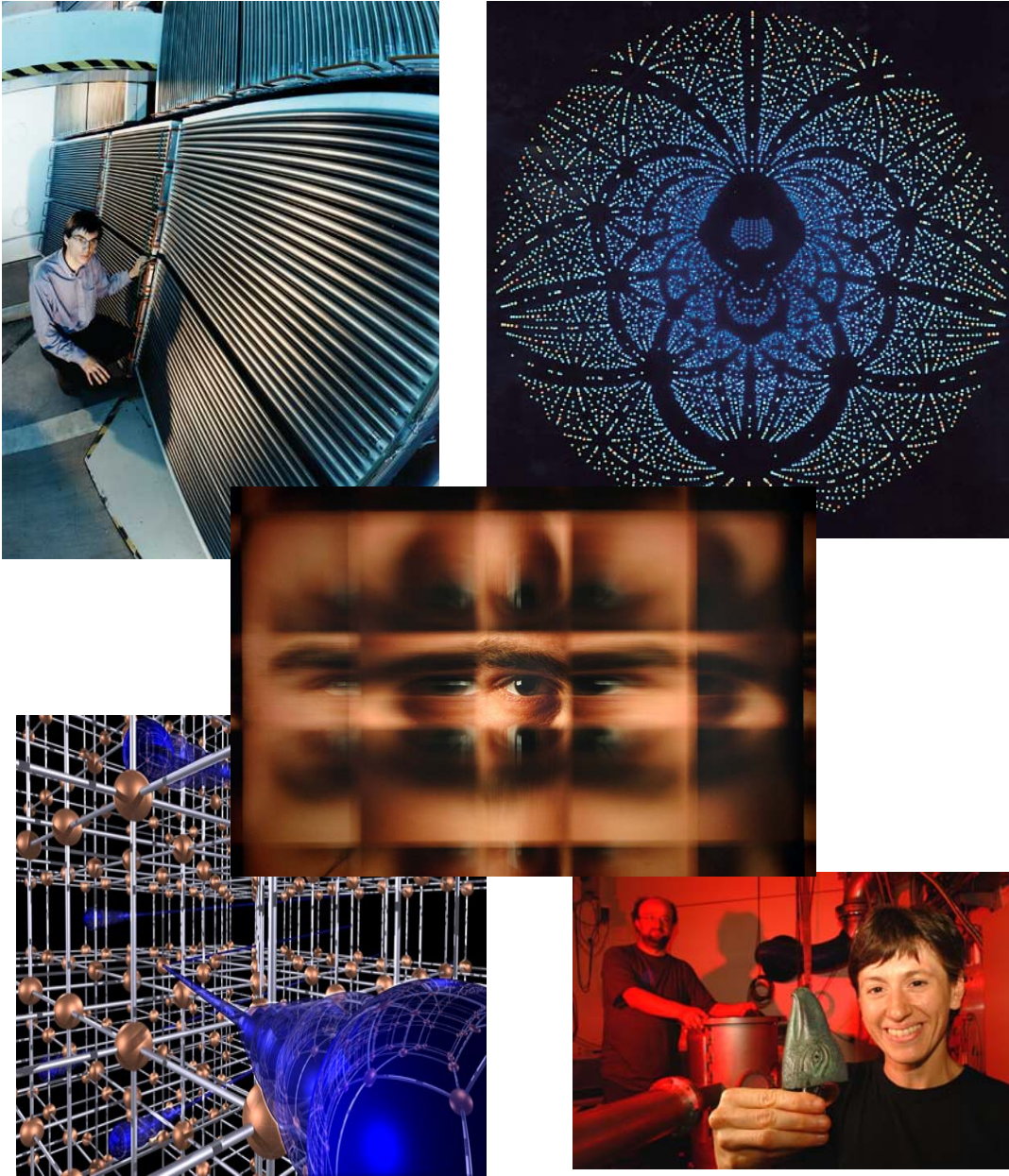


20/20 Vision: an e-Infrastructure for the next decade



Report of the Data and Information Creation Working Group to the e-Infrastructure Steering Group

20/20 Vision: an e-Infrastructure for the next decade

**Juan Bicarregui
Richard Boulderstone
Lorraine Estelle
Jeremy Frey
Neil Jacobs
William Kilbride
Brian Matthews
Robert McGreevy**

31 March 2006

Summary

We cover 5 aspects of data creation:

1. The nature of the data itself
2. The creation of data by physical research
3. The creation of data by e-research
4. The creation of data by digitization or repurposing from other sources
5. The integration and certification of data

Throughout the document by data we mean information and data.

For each topic we give current issues and discuss the functionality to be provided by the future e-infrastructure. We then give outline of some new aspects of the programme of work required to achieve it. We conclude with some general comments on implementing the programme work and some “don’t needs” .

Summary of current position and “future ideal” Infrastructure

We discuss the current position and some features of the future “theoretical ideal” e-Infrastructure provision. For each of the five topics we highlight some issues and give a vision of how they will be managed by the future e-infrastructure. We then discuss the topic in general

The introduction gives a high-level view of the expected benefits arising from investment in the e-infrastructure, the next 5 section cover each of the 5 topics in turn.

Throughout this document, we use data to mean data and other research output information such as papers, documents, software, etc.

Introduction: The benefits of an effective e-Infrastructure

We expect the future e-infrastructure to accelerate progress by enabling both better research (in the efficiency and quality) and new avenues for research (in opening up greater opportunity for cross disciplinary research).

Better research

- **Openness** – the e-infrastructure should support a culture where research information is available as widely as possible whilst respecting commercial or other reasons for restricting access. This will yield greater effectiveness and efficiency in the scientific process by facilitating wide collaboration and will foster transparency, auditability and the widest possible cross disciplinary opportunities. However, openness also places considerable demands on the infrastructure to provide mechanisms for the protection of intellectual property which engender sufficient confidence in users for them to employ its mechanisms for security, provenance, and appropriate distribution rather than to use ad hoc, overly restrictive means to protect their interests.

New research

- **Integration** – the e-infrastructure should facilitate the understanding of research information by as wide a group (of people and tools) as possible in order to support the advancement of research of a multidisciplinary nature. A common semantic classification of information from all disciplines will minimise the cost of integration of resources and therefore support improved cross disciplinary research (even enable new cross disciplinary research which would not have otherwise been economically feasible). (Who can tell what possibilities may be opened up in the longer term future?)

The report identifies 5 topics and for each one describes the current provision and some features of a theoretical ideal infrastructure. It then describes the programme of work for each topic. It concludes with list of general issues and “don’t needs”. The 5 topics are:

1. The nature of the data itself
2. The creation of data by physical research
3. The creation of data by e-research
4. The creation of data by digitization or repurposing from other sources
5. The integration and certification of data

Topic 1. The data itself



The future e-infrastructure will more directly support the management of data throughout its lifecycle “from cradle to grave”.

Issues

Data identity. Persistent Unique Identifiers (or an alternative means to achieve this functionality) will enable global cross referencing between data objects. Such Identifiers will not only be used for data and software but also for other resources such as people, equipment, organisations etc.

On the other hand, any scheme of identification is likely to undergo evolution so preservation, and in particular integration of archival and current data, is likely to require active management of identifiers.

Data objects. Data will be made available with all the necessary metadata to enable reuse. From its creation, and throughout its lifecycle, data will be packaged with its metadata which will progressively accrue information regarding its history throughout its evolution.

Data agents. Data will be “intelligent” in that it maintains for itself knowledge of where it has been used as well as what it uses. (This can be achieved by bidirectional links between data and its uses or by making associations between data themselves stand alone entities. In either case, active maintenance of the associations is required.)

Software. Software will join simulations, data, multimedia and text as a core research output. It will therefore require similar treatment in terms of metadata schemas, metadata creation and propagation (including context of software development and use, resilience, versioning and rights).

Data Forge”. Rather like sourceForge for software. We imagine a global (probably distributed) self service repository of the for data which is made available under a variety of access agreements.

Discussion

By 2020, “Data networks will have gone from being the repositories of science to its starting point”¹. Although this quotation perhaps overstates the case, it is clear from current trends that the creation of new knowledge by synthesis of data from existing or ongoing experiments will become increasingly important in the next decades. However, behind this vision, lies a requirement on the data management technology for far greater ease in data collection, interoperation, aggregation and access. Technology and tooling must be developed to meet this requirements both in the manifestation of the data itself and the software that manages it. Critical to this will be the collection, management and propagation of metadata along with the data itself. Technology and tools must be developed which facilitate:

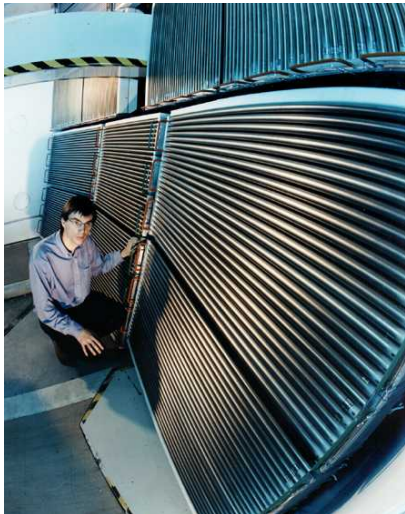
1. Metadata schema creation with optimal reuse of existing schema hence maximising the potential for data interoperation,
2. Metadata creation as early as possible and as automatically as possible, and
3. Metadata propagation, so that data and metadata are managed together by applications ^{hence} enabling additional types of analysis of the data at each stage in its evolution (provenance, audit, etc).

A simple example of first of these would be tools which use standardised domain ontologies to find the “nearest existing” schema when attempting to define a new one. The second point would be supported by a highly instrumented environment for research with context aware knowledge of the processes being enacted. The third requires data analysis and transformation tools to be enhanced to maintain metadata relations and an infrastructure which enables data and metadata to be handled “as one”. One mechanism which could support this would be the packaging of data and metadata into “data objects” which could self-maintain for example by managing bi-directional links to related data objects which either use or are used by this data. A simple example of a data object could a person specification which, along with name strings etc., self-manages its links to current affiliation, other people etc. Data and tools using the person object would then be liberated from needing to maintain these associated fields.

Along with these technologies will have to be developed the “library” of the data objects themselves. There will be, as ever, a pull-push relationship between the development of the technology and its use in establishing the datasets which it manages. For the person objects example above, one mechanism which could kick-start the creation of this data pool would be a number of pilot projects which would cover a large proportion of the UK researchers based around existing “hubs” which already have large numbers of researchers on their books.

¹ [Declan Butler, Nature New Feature “2020 computing: Everything, everywhere”
<http://www.nature.com/news/2006/060320/full/440402a.html> doi: 10.1038/440402a]

Topic 2. Data created by physical research



The future e-infrastructure will reduce the cycle time from conducting research, through analysis, publication and feedback into new research.

This section covers data created from physical research in the laboratory, field, or facility.

Issues

Metadata. Key to reuse of data is the (automated) collection of metadata – as much as possible and as early as possible. The ideal is that all the information about an experiment, the environment, the people as well as the data flowing from the equipment and samples is captured in a digital manner as early on in the process as possible.

Research plans. For hypothesis driven research, the research plan will provide a digital context to form the basis of automated capture of data and metadata and the policy for data distribution. For exploratory research, the analysis tools will provide some assistance in gathering of metadata.

Real-time analysis of data enables new possibilities for efficiency such as a “just enough” approach to experimental planning which limits data collection to just sufficient to prove or disprove the hypothesis thereby releasing researchers and other resources to move on to the next experiment more quickly. This type of “Heads up” process management (perhaps supported by expert systems technology) should be used to optimise the efficiency of the research process by dynamically modifying the experiment as partial results become available. (Note that the modelling involved is likely to be non-linear and that statistical analysis will be required as well as application specific data analysis.)

Data Acquisition Tools . Many instruments come with manufacturer specific data capture tools. Pressure is increasing on the manufactures to conform to standards at least in the way the data collected can be exported from their software - the XML trend is increasing. If this can be pushed further on to use Semantic Web/Grid technology, to

include the proper semantics and unique identifiers (use of ontologies when possible, RDF etc) then data integration will be come much easier. When several machines need to be controlled and orchestrated, then tools like LabView, HPV, which are visual programming languages, are becoming increasingly used. However, there use in more distributed environments, is still largely uncharted. Annotation at source is the ideal.

Data Analysis Software This is often subject and experiment specific. Software packages such as IGOR, R, even EXCEL, plotting packages, need to inter-operate in a more robust manner and the ability to keep track of the workflow, provenance etc is crucial.

Electronic Lab Notebooks. ELN is a rapidly growing area. True ELN's for use in the laboratory or in the field, rather than for 'after-the-fact' write up, are less common. Industry is concerned with information capture, re-use, enforcement and legal issues. Academic labs are more concerned with versatility. The Human Computer Interfaces of ELNs are very important. The resulting information must be available in different contexts and for different purposes (the lab, the office, for the manager, for the future researcher etc).

Virtual Access to Physical Facilities Virtual access to physical facilities will become increasingly important as larger teams are needed in multi-disciplinary research which can't all be physically present at an experiment. There is a push to provide remote access to equipment and the resulting data. There is also a need for provision of other services to support the data. All this functionality and service provision needs to be content aware.

Discussion

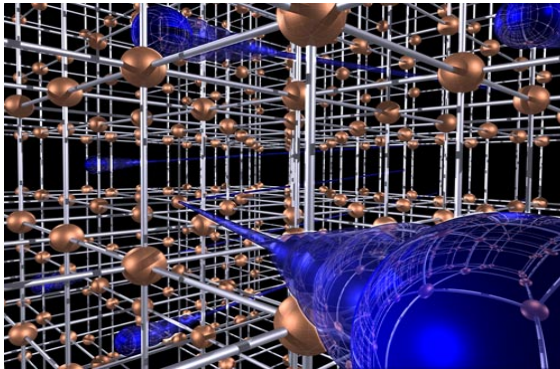
In the context of scientific discovery, the laboratory and field notebook has for several centuries been an essential tool of the scientific method. This methodology (document planning, experiments, results, thoughts etc.) continues and will drive high quality, reproducible science but, by 2020, the paper notebook will be replaced by the electronic, virtual notebook. The efficient use and re-use of research data will therefore be even more highly dependent on the quality of the metadata associated with the data. It will be vital to transfer the data to a digital form as early as possible in its creation, to collect the associated context at the same time, and to ensure that this information is propagated together along the research chain. This will enable much metadata to be captured automatically, and transparently and associated with other events (such as weather conditions, environmental aspects of the laboratory, people) in a seamless manner. The ability to switch the context of the lab notebook, and obtain different views depending on circumstances, further enhances the flexibility of the electronic version over the paper notebooks.

Viewed in this wider context, the ability to capture information about the environment of an experiment, is further enhanced by the pervasive aspects of modern computing, an areas which is growing extremely rapidly. Research can take advantage of the rapidly falling cost of sensors to ensure that experiments can be monitored in a far more automatics and extensive manner than has been possible in the past, creating a much richer source of data for subsequent study. As we push towards a smaller, greener,

footprint, the modern 'smart' buildings will deliver a much greater built in sensor & control network to the laboratory, simply as part of the infrastructure of the building.

The quantity of data and metadata captured by this approach will be a significant increase on current data flows, which are already rising exponentially. The quality of data collected in this manner will mark it out as much more useful in the longer term. However, it will be important to store the information carefully. This is a role for a hierarchy of storage systems corresponding to the different administrative domains active in the research environment (researcher, research group, department, university, national collection, journal, private database holders etc.). These ideas are currently being addressed by institutional and subject based repositories. These repositories will grow in size and depth over the next 10 years and enable researchers to access material by tracing a path from published literature all the way back to the laboratory notebook, all via a digital pathway.

Topic 3. Data created by e-research



There will be a much greater use of e-research and much tighter integration of physical research with e-research.

Issues

New opportunities. e-research, that is research conducted entirely on computers, such as in simulations or ab initio calculations, is playing an increasingly important role in the process of scientific discovery. As well as opening up new avenues of research it can also provide a reduction in cost of several orders of magnitude and so open up the possibility of a vastly increased number of experiments. With these increases in scale come requirements for improved reliability, data analysis and statistical analysis software, etc.

Auditability. There will be however be complex issues surrounding, accuracy, reproducibility and audit, a process akin to sensitivity analysis of the research outputs. In e-research, just as in physical research, there will be a need to automatically “understand” and make use of previous data. An e-research equivalent of the lab notebook culture will need to be developed.

Integration of e-research and physical research. Computation and simulation in real-time will increasing become a integral part of the process of undertaking physical experiments with e-results, being used to drive the collection and analysis of physical processes. This will enable much more effective and efficient use of physical resources.

Large scale research. The e-infrastructure will also be required to support the large scale data processing and analysis driven by large scale e-experiments and large scale, community based research involving hundreds or thousands of researchers linked by webs & grids. The possibility of conducting such large scale research will open up new avenues for research in many disciplines.

Discussion

There is considerable scope for improvement in research efficiency, and indeed in research capability, by closer integration of physical and simulation based research. At present, e-research, such as computer simulation and modelling, and experimental

research tend to pursue parallel tracks, with comparison of results at the end of each process. There are many potential benefits to be brought about by tighter integration of these two processes, for example, by real-time feedback between experiment and simulation.

In general, these advances can be achieved by the efficient deployment of existing technologies, together with effective integration technologies, rather than requiring new technological developments. For example, the real-time feedback mentioned above may require the use of 'HPC' capability specifically at a particular time and for a defined period. This type of requirement is well suited to Grid systems as long as appropriate real-time scheduling systems can be suitably developed.

The results of e-research should be considered as data in exactly the same way as for experimental research, with the sole exception that in principle e-research results are exactly reproducible if the relevant software and all necessary input data/metadata are available. This will raise the question of whether it is more efficient to curate the complete outputs of e-research studies, or to only curate the inputs plus selected outputs. This will be a subject specific decision which will depend on the cost of, and possibility for, easy recreation of the outputs, compared to the cost of storage. Where complete outputs are curated it is of course necessary that the inputs (which in any case tend to be smaller) are also curated.

One type of on-line feedback is where real time (but compute intensive) analysis is used to produce the 'final results' as the experiment proceeds. Experiments can then either be continued, stopped or modified depending on the actual results, rather than a pre-ordained plan.

Another type is where the experimental results are used to incrementally modify a pre-calculated model. The measurement + modelling then only need to be continued until convergence is achieved, or can be stopped if it is clear that this is not going to happen. A specific example of this (which happens to have major industrial relevance) would be the use of Finite Element Modelling (FEM) in relation to experimental measurements of residual stress in e.g. aircraft components. FEM is a commonly used modelling technique, but typically it only provides qualitative rather than quantitative results. However, it is cheap in relation to an experimental measurement of corresponding scale and resolution. The initial FEM model can guide the measurement to the key points. The data at these points can then be used to constrain a new model, which then provides a new set of key points for measurement, and so on. Parameters from the final FEM model will improve the results of future of similar materials/components, so there are considerable potential efficiency gains. (But clearly here this requires the existence of accessible results databases, to ensure the forward propagation of the knowledge obtained.)

e-research techniques also provide the possibility for new types of experiment, through for example the integration of enormous numbers of measurements/observations in a single experiment (e.g. sensor inputs used for modern climate modelling), both in a structured or unstructured way (e.g. community based observational inputs via the internet). These increases in scale also bring a requirement for improved reliability (of the internets or grids more than the sensors), new data analysis software etc.

2.4 Data created by digitisation and repurposing



The future e-Infrastructure will support the use for research purposes of data collected for other purposes.

Issues

Strategic Investment. Digitisation is not an end in itself, it is a means to an end, and whilst the last decade has seen a rush to digitise within the UK public sector, supported by considerable sums of money², as well as providing digital content, experience gained on these digitisation programmes has helped build capacity with ICT and has amplified strategic weakness such as digital preservation needs³.

Repurposing of data for research. Increasingly scientific research will be able to be done by harvesting data that already exists for other purposes. Data standards across government departments will enable large-scale research particularly in sociological and medical disciplines. New research methods will also be enabled by the e-infrastructure and its ability to co-ordinate large number of technologically equipped “amateur” scientists. Community supported research conducted by interest groups such as environmental or orthinological organisations will become increasingly important and valid. Data collected by the private sector (for example from loyalty cards) will also be available for scientific as well as commercial marketing purposes.

Public engagement. The e-infrastructure will also support initiatives promoting the public engagement of science. In this respect openness pays dividends in the long-term through public engagement, an additional benefit over and above direct commercial exploitation of research results. New community-based research methods may require

² Early programmes such as the Research Support Libraries Programme, JISC Fast-track Digitisation and New Opportunities Fund Digitisation programmes have been succeeded by more recent initiatives such as the JISC Digitisation Programme. These national endeavours have been replicated at a local level (see for example the Local Heritage Initiative <http://www.lhi.org.uk/>), and with large scale international collaborative programmes (for example the E-content Plus programme http://europa.eu.int/information_society/activities/econtentplus) The UK is not alone in this drive (see for example the National Digital Archive Project, Republic of China <http://www.ndap.org.tw/>)

³ [Waller and Sharpe 2006, Mind the Gap, Assessing digital preservation needs in the UK, Digital Preservation Coalition, 2006.]

coverage of infrastructure beyond that required otherwise, for example beyond the universities to the household.

Discussion

Different stereotypes of users reasons for choosing digitised resources have also emerged and five example use cases are delineated below. It is noted that one way to maximise the use of a resource is to maximise the number of ways in which it may be deployed. Consequently, maximising the re-use of a digital object means identifying, reducing and where possible eliminating those barriers. The use cases are:

- The digital surrogate
- The digital reference collection
- The digital research archive
- The digital learning object
- The digital archive

These use cases are neither self-contained nor exhaustive and many resources could fit into more than one category.

Digital surrogates provide access to objects that cannot be easily reproduced or distributed. Classically this has been used to refer to rare or precious cultural objects (for example the early Beowulf Manuscript in the British Library), but it might reasonably also include any digital surrogates that let researchers carry out computer-based analysis. Consequently the outputs of scientific instruments and a range of detectors might be termed digital surrogates. The use case for digital surrogates is for the group of experts that cannot access the original resource but is able to use the surrogate to undertake significant parts of their research.

A **digital reference collection** has two likely uses: as a finding aid for other resources and as an authority upon which analysis or discussion can be built. The former might be characterised as a metadata service of one kind or another and is often associated with a library or bibliographic function while the latter tends to be more analytical and is specific to individual research communities. However, simply creating a thorough catalogue of a particular type of resource can be a work of considerable scholarship, so the two functions overlap in production and use. The use case for such digital reference resource is the desk-based assessment at the start of a research project or as an authoritative source throughout the project.

The **digital research archive** is the digital residue that derives from research. A research project may quite legitimately have an analogue output as the principal goal of its research: classically a book, article or monograph. Although the project does not set itself the goal of creating digital objects per se, but may nonetheless created digital objects to support an analytical function. These digital objects may present further opportunities for research: either to test and revise the conclusions drawn in the initial project through re-analysis; or to provide a knowledge base to which additional data may be submitted. The use case for such archives exist within the research process with an expectation that significant findings should be constantly assessed and revised.

Digital learning objects have seen particular expenditure over the last ten years and the literature on them is extensive. It is sufficient to say that these may be assembled

from many sources, including all of those listed above. However the creation of digital learning objects has historically been a major driver in the digitisation process (for example the TLTP programme), and many digital objects are created purely for the purpose of education. The immediate use case for a digital learning object is the teaching and learning context in which it is deployed but arguably any component could be used more widely than the learning context.

The **digital archive** consists of the digital objects associated with an event or process. Though not inherently created for research the digital archive may become of interest to academic research at some point in the future, and may also have legal or political implications too. This is a superset of the digital research archive, insofar as it is a by-product of some other activity. The use case for this type of resource is deferred and thus unpredictable.

The above illustration of use cases helps make a broader point – that maximising the use of digital resources for research is not simply a question of access agreements, cataloguing and promotion. The nature of digitisation can have a profound impact on the sorts of use that a resource will sustain and maximising return on the means extending the use cases that a resource can support.

The clearest example of this phenomenon is in regards to learning objects. As long ago as 1995 Laurillard noted that we should not confuse access to digital resources with teaching or learning. Selection, pathways and guidance into or through the digital resource, and tasks that help students explore aspects of the resource are essential. Consequently much of the promise of open-ended problem based learning with digital resources has proven to be illusory. Access is not education. Research archives can be educational but other activities are needed to make them so.

The same is true of other use cases. If the intention is to present a digital surrogate then the digital object has to be at least as accessible as the original: but for a digital archive security and trust may be of greater importance than access. A reference resource is likely to require mechanisms to ensure currency and accuracy: a digital research archive would be inclined to prohibit subsequent corrections and ensure that any subsequent additions are monitored and clearly signalled. Research archives will need to provide comprehensive access to the original data so that meticulous researchers can evaluate its entire contents: a reference resource will be optimised to allow search and retrieval of discrete chunks of information.

The method and goal of a digitisation project can constrain or enable the development of such additional use cases. An image taken for educational purposes could conceivably be reused as a digital surrogate. A digital surrogate could feed an educational resource, a reference collection or a research archive. These categories of use case are not self-contained, so any single resource could reasonably be used in each of the above contexts. The resource would simply need to be configured and made accessible for these different types of use. However the opposite also applies. Poor design or insufficient metadata capture at digitisation may prevent re-use in research; lack of attention to preservation may result in data loss; difficult to use formats may inhibit educational use.

Digitising has many benefits, but digitisation is not an end in itself, nor should users be taken for granted. Evaluating the uses that a resource can support, and understanding how those uses can be expanded is critical to the success of any digitisation project.

2.5 Data integration and certification



The future e-infrastructure will be based upon standards which support uniform classification, integration, certification and citation of data across all sources.

Issues

Standards. Central to much of the vision above is an assumption that interoperability will be greatly improved through the use of standards of many forms. Equipment standards, Software standards, Information standards, will all be in place to enable sharing of data on a global scale. To engender this, both technical and behavioural aspects related to the development and use of standards will need to be addressed.

On the technical side, a steadily increasing proportion of the instrumentation used in experimental research, i.e. in experimental data creation, is commercially manufactured and is often well advanced in the application of new e-techniques. However the data standards used are usually manufacturer specific, even for similar types of instrumentation (as epitomised by the case of the two next generation DVD formats). The public sector has a clear role to play in developing and encouraging common standards, particularly as the future demand will be increasingly for the ability to combine data from a wide range of instrumentation.

On the behavioural side, the research culture must evolve to make it advantageous to the individual researcher to follow standards? Credit needs to be given for more than publications. Issues around an ownership should not limit openness. "What's in it of me?" needs to be addressed.

Classification. Information standards will be based around widely used and commonly agreed ontologies, thesauri, controlled vocabularies, etc. Mechanisms will be in place to bring together ontologies (and other information standards) from different disciplines. A place to lodge standards, a managed repository of standards, covering technical standards, quality standards, provenance standards, will be available and in use by all disciplines.

Open access. The above vision relies on minimising the barriers to access to research outputs. The e-infrastructure will need to support a variety of business models if the

content is to me made as accessible as possible with limits to access only for good reasons. This approach should apply equally to both existing and new research, and be irrespective of scale.

Data citation. The infrastructure will support a system of academic review and credit for use and citation of all forms of research outputs, data as well as publications, negative as well as positive results, licenses patents, software, etc.

e-learning. As well as enabling information integration, Ontologies will also be used as a means for learning. For example in classifying a piece of data as an air temperature, the ontology will ask for altitude, humidity, etc?

Software integration (OLE for science). Scientific software will be interoperable so that outputs from different tools can be “cut and pasted” in a functionally enabled way (like OLE does for MS Office tools). Thus a graphs included in a paper will be clickable to access not only the underlying analysed data, but also analysis software that produced it and the data which went into that, etc.

Propagating best practice. The e-infrastructure will support an *information infrastructure* which will provide a means to propagate good practice across disciplines. Up to date information on research methods, teaching, legalistic and ethical practices, management of ownership, IPR, copyright and privacy, quality assurance, integrity, authenticity, trustworthiness, provenance, audit and accounting will all be available on specific projects and in summary form. This data will be employed in the reviewing of projects and programmes. There will need to be a balance struck between Flexibility and auditability. Health, Safety and Environment issues require auditability. As do validation of results. On the other hand, these issues should not limit the scientists flexibility to dynamically change the direction of their investigations and their research plan during the research process.

Discussion

An essential first step for the exchange of data and information is the appropriate means for data storage and access. The funding and development of robust institutional repositories and thematic repositories will allow for the storage and access of data and other research output. To date, much of the work in the field of institutional repository work has concentrated on research publications, but further work is being undertaken to encourage the growth of repositories of other forms of research outputs.

Critical for the exchange and sharing of data is not only the infrastructure for storage but also the cultural change which will reward the deposit of data in appropriate repositories and recognise the impact of this data. Currently there is no universal standard for citing data. Indeed, ‘citing unpublished data (data that has not been published in the scientific literature) in the references, even when electronically archived and made available through the Internet, is not allowed by some journals.

Collaboration with scholarly publishers is an essential in promoting the cultural change necessary to encourage the citation of data. Publishers could provide authors with the guidelines and assistance in the preparation of citations of data and other information.

Data repositories must also be equipped with the tools that will facilitate ease of citation. When citing data, researchers will require rich and readily accessible metadata that will provide accurate information about the authors and developers of such datasets and the location of the data.

In the world of scholarly communications researchers are recognised through citation. The commercial sector provides the index citation databases that offer a valuable measure of the formal scholarly impact of articles, journals, and authors. Not only do such databases provide a measure of impact, they provide a means of resource discovery – enabling users search and find relevant articles.

It is interesting to note that the commercial sector is already developing citation databases of scholarly content from open access institutional and subject-based repositories. Repositories of such content adopt the information standards that allow for 'machine to machine' harvesting of metadata are being searched and cited by these emerging services. The emergence of these commercially provided citation indexes will provide effective resource discovery and also measure the impact of research outputs deposited in such repositories. However, as yet there is as yet no equivalent measure of citation or impact of data.

The commercial sector is likely to provide such a service if there is an enabling infrastructure and standards. This will require data deposit in open access repositories with the metadata that can be harvested so that it can be included in commercially provided services. Such developments will themselves reinforce the cultural change required to further encourage the deposit and citation of datasets.

Datasets of positive results are not the only output of research that can be deposited and made accessible: Software is a research output and a similar programme of infrastructure, standards and rewards are required to encourage deposit and citation.

The publication of negative results should also be facilitated. The publication of such results can avoid the repetition of costly experiments. The availability of negative results has also the potential for use other than the original purpose. For such economy of use it is essential that the data collected in negative results is collected and deposited in accessible and open repositories.