ARCADIA /CAMBRIDGE UNIVERSITY LIBRARY

An investigation of how researchers in data intensive scientific fields use, process and curate data

An Interim Report

Yvonne Nobis

11/1/2011

Contents

An investigation of how researchers in data intensive scientific fields use, process and curate data

Executive Summary

This project was an Arcadia funded project, examining (i) whether research outputs from data-rich science (most typically software) were failing to be disseminated, both within and outside subject communities and (ii) whether there was a useful role for information specialists in this area.

Three distinct subject domains were investigated: genomics, structural biology and physics. This Interim Report covers only the first two fields. The report on Physics is currently being completed.

Key Findings

- Large collaborations proved many examples of best practice, both in the physical and biological sciences. There was not perceived to be a real problem at this level.
- There is an acknowledged problem at the level of smaller projects across all domains.
- Barriers to good practice include cultural norms, isolationism, working practises, lack of training in research data/software management, funding and institutional goals focused on core areas, and lack of a mechanism for effective communication with colleagues outside the domain.

Is there a role for the information professional?

The investigation suggests that there is in the following areas:

- Facilitation of workflows
- Integration of new preservation and cataloguing paradigms
- Training researchers in data management and workflows
- Provision of a service to curate software from research groups and to help make this discoverable.

Background

In an article entitled "The Data Big Bang and the Expanding Digital Universe: High – Dimensional, Complex and Massive Data Sets in an Inflationary Epoch[1]", Pesenson *et al* set out many of the problems facing astronomers in dealing with massive, intrinsically complex datasets. They argue that "data intensive astrophysics requires an interdisciplinary approach" and that the richness and complexity of the new datasets can only be analysed by new sophisticated tools. They call for new paradigms for the analysis and visualisation of such data and suggest solutions from the fields of applied mathematics, statistics and artificial intelligence.

This increase in complexity and scale of datasets, and the techniques needed to analyse data, is equally true of most other areas of scientific pursuit.

Novel Science demands novel solutions. Yet how are the processes of such analysis being recorded? How (unless the argument is expressly promulgated) is a researcher in one field (for example, astronomy) to realise that techniques from another (applied mathematics) may be potentially useful?

The project set out to query how the techniques employed for data analysis from one discipline can be applied to another, and whether there is a role for the librarian in assisting this process.

Such re-use of techniques employed in analysing data is occurring but in only in a limited capacity. An example is PathGrid: collaboration between astronomers at Cambridge's Institute of Astronomy and Cancer Research UK's Cambridge Research Institute where image-analysis software developed for astronomy is being used to automate the study of pathology slides. [2] The author was interested to find out how this project was conceived and what the drivers to initiate such an inter-disciplinary project were. Could this success be replicated?

Issues discussed include those at the macro level of research council mandates, funding streams, discipline cultures and organisational drivers and publishing models as well as at the micro institutional level: whether the support and training provided to researchers is adequate.

The emphasis in this paper primarily relates to the curation of data techniques used on 'live' data (as opposed to that of archived resources, arguably the traditional domain of the librarians). Inevitably many of these issues are linked to the accessibility and preservation strategies relating to the data itself.

To date, most information science/library-based research on the data outputs of scientific research has concentrated upon these issues in relation to the storage of, and access to scientific data, and the preservation issues arising therefrom.

The data problem in context

A recent editorial (February 2011) in the journal *Science* highlighted several problems involved in dealing with data management and storage. The article argued that the point has been passed *where more data is being collected than we can physically store*.

The scale of the problem has often been described and is in flux (and indeed will have significantly increased even during the course of this project).

Hilbert *et al*[3] attempted to estimate the word's technological capacity to store, communicate and compute information during 1986 to 2007. They concluded that the world's technological information processing capacities are now growing at an exponential rate, and that application of Moore's law (that the number of transistors on an integrated circuit double approximately every two years) gives surprising results: that the per capita capacity of the world's general purpose computers has doubled every 18 months, however comparable storage capacity per capita has doubled only every 40 months.

Comment [JN1]: Strictly speaking, Moore's Law stipulates a doubling every 18 months/

Comment [y2]: I defer: should I leave it in for now, or find another example , which I know I have come across?

Librarians tend to enjoy calibrations of scale relative to the amount of data held in the US Library of Congress. In those terms, one high-throughput DNA sequencing machine can read approximately 26 billion characters of the human genetic code: 9 terabytes of data in a year: alongside the related information generated this equates to 20 new US Libraries of Congress each year.

Comment [JN3]: Wow!

The issues researchers face as a result of the data deluge as reported by *Science* (who had 1700 respondents to an online questionnaire on the data issues) include "lack of common metadata and archives" and that this was the "main impediment to using and storing data… most of the respondents have no funding to support archiving". This was found to be equally applicable to disciplines with well-established data archives, such as genomics. Some 20% of the respondents to the *Science* questionnaire regularly use or analyse data sets exceeding 100 gigabytes, and 7% use data sets exceeding 1 terabyte. Of the respondents to the *Science* survey nearly half stored their data solely in their laboratories.

The *Final report of the High level Expert Group on Scientific Data; Riding the wave: How Europe can gain from the rising tide of scientific data: A submission to the European Commission (October 2010)*[4] defines a fundamental character of our age as the 'rising tide of data' and proposes that we are on verge of a great new leap in scientific capability, fuelled by data. This point is also made in Hey, Tansley and Tolle (Eds), The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, 2009.

Comment [JN4]: This point is also made in Hey, Tansley and Tolle (Eds), The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, 2009.

Comment [y5]: Added this in as good Background reading and the only monograph referred to

This recognises the importance of data as an asset in itself, and the problems of data-intensive science operating at distances where many of the protagonists have never met, or indeed communicated. Such data requires "professional analysis and engineering", and achieving an interoperable system in the midst of such (data) heterogeneity is a significant challenge. "Researchers and practitioners from any discipline should be able to find, access and process the data that they need…cross fertilisation of ideas and disciplines will produce novel solutions and promote greater understanding of complex problems".

As part of the outcome of this research a scientific e-infrastructure 'wish list' was created, and challenges to overcome to attain this are also listed. These include assurances that data is collected with the information necessary to re-use it, issues of trust, usability, interoperability, reward structures, and preservation and sustainability.

Comment [JN6]: "are"? (data is plural) But it's amatter of preference and switching would create lots of knock-on changes, so perhaps best leave it as it is.

Alongside proposing the development of an international framework for a Collaborative Data Infrastructure the high level committee queried "how can we foster the training of more data scientist and data librarians, as professions in their own right?

In a similar vein, the *Baseline report on Drivers and Barriers in Data Sharing*[5] (a FP7 project[1]) was published in October 2011.[6] This makes the point that the potential of e-science can only be met by adding an interoperable data sharing, re-use and preservation layer on top of the emerging connectivity and computational layers. This Baseline report aims to identify, collate, interpret and deliver evidence of emerging best practise in sharing, re-using, preserving and citing data (this is a long term project due to report in 2020). Interestingly libraries are put at the core of their activities: "step by step libraries, data centres and other infrastructure units are intensifying their activities in the fields of research data management". An initial series of interviews,conducted by the authors of the report with relevant stakeholders produced 14 different categories or perspectives which are barriers or drivers to data sharing, and these will be used in the next phase of the project to examine researchers attitudes to data sharing.

These categories include many of the following : education (data sharing needs skills), behaviour (it must be easy to share data), Incentives, appreciation and recognition, funding, legislation and culture/attitude, amongst others.

Authors from both of these reports were interviewed in the course of the research for this Arcadia project.

The Royal Society has undertaken a major project, *Science as Public Enterprise*[7], which aims to identify the "principles, opportunities and problems of sharing and disclosing scientific information" and how such information should be managed to support the open exchange of data and ideas, both to other scientists and to the public. This group has yet to report.

However for the promise of scientific data to be fulfilled by scientific discovery, the nature of the discourse between researchers both within and outside their own disciplines, and with their research support services, has to keep pace with the changing data eco-system. As Wolpert et al[8] state "many of the habits of scientists have barely changed since the 18th century. Driven by curiosity, they have typically pursued their research, published their findings, usually in peer-reviewed journals, filed their data, and then moved on".

The problems inherent in data management and the strategies adopted have been the focus of much research. "What researchers want"[9], a comprehensive review of literature in the

Comment [JN7]: 'queried' (otherwise sentence isn't right.

Comment [JN8]: What's FP7?

Comment [JN9]: Conducted by whom?

Comment [JN10]: Footnote? http://royalsociety.org/policy/projects/science-public-enterprise/

[11] 'Framework programmes' are the main financial tools through which the European Union supports research and development activities covering almost all scientific disciplines. FPs are proposed by the European Commission and adopted by Council and the European Parliament following a co-decision procedure. The Seventh Framework Programme (FP7**)** bundles all research-related EU initiatives together under a common roof playing a crucial role in reaching the goals of growth, competitiveness and employment.

area, drew the conclusion that "it makes sense to invest in better data management during the research phase because doing so will improve data preservation once the research phase has ended". Several other studies have looked at potential methodologies for handling data and whether these should be made mandatory by the Research Councils. [10]

The 2008 RIN study, *To share or not to share : Publishing research data*, (RIN) a jointly funded study with JISC (Joint Information Systems Committee) and NERC Natural Environment Research Council, was a comprehensive survey into how researchers were responding to funder mandates to make their data more available and accessible to others. Over 100 researchers were interviewed.

The study found that there are significant variations – as well as commonalties – in researcher's attitudes, behaviours and needs, in the available infrastructure and in the nature and effect of policy initiatives in different discipline and subject areas.

Research conducted for this project bears this out. Many datasets and the techniques used in analysing the data were felt to be of potential value to other researches and users- particularly those arising from small scale projects,. The RIN survey found that these datasets tended to be not managed effectively or made readily accessible and re-usable.

Comment [JN11]: To what does 'these' refer?

Many research funders are putting policies in place to ensure that datasets judged to be potentially useful to others are curated in ways that enable discovery, access and re-use to optimise the value and uses of data produced during the course of the research that they fund[11]. The RIN authors found that there was not a perfect match between those policies and the norms and practices of researchers in a number of research disciplines.

The report concludes that there were two reasons for making research data publicly available: to make them part of the scholarly record is that they can be validated and tested; and to enable them to be reused by others in new research.

The emphasis of this Arcadia project was is slightly different: it is not concerned with the curation of data *per se*, but *with the curation of software techniques used on to analyse the data, and the curation and discoverability of these, both within and outside disciplines*. Arguably this can also be thought of as a 'data product'. Are researchers in one area merely replicating techniques commonly used elsewhere? If so, how can and how should this be addressed and what is the role of the information professional?

The particular problem of software code

Increasingly the ability to write software is an intrinsic part of scientific endeavour, and is crucial when large datasets are being manipulated, analysed and the results interpreted.

In a commentary on the 'climategate' debacle Ince[12] observes that "if you are publishing research articles that use computer programs, if you want to claim that you are engaging in science, and the programs are in your possession and you will not release them, then you are not a scientist" . This may be an extreme view but it does highlight the problem.

The data life-cycle[13] within the research process has four distinct stages: the collection of experimental data, the cleaning of this data, the working on the data and the curation of the data.

The authors of *Data in Motion: a new paradigm in Research Data Lifecycles*[14] put the case compellingly.

"An increasing amount of data managed as part of the modern scientific discovery process is derived data generated by computational and analytical methods. The tools used to generate this derived data, as well as the versions and configurations of these tools at the time this data was produced, must be preserved along with the datasets. Absent this requirement, the data and any errors which might have been introduced cannot be fully understood and/or replicated. This requirement should also extend to observational and experimental data which is collected and stored by the use of computational tools for the same reasons."

Most often data that is made available is not raw data but derived or redacted data and the processes on this are usually recorded (and this process is also not without its problems - as interviews with some members of the astronomical community will attest).

De Roure and Goble[15], are computer scientists who work closely with bioinformaticians. They have set out guidelines for good software practice for scientists and make this point expressly: "Science", they write, "is becoming increasingly digital, and scientists' tools are not just the experimental apparatus of the laboratory, but also the software apparatus that they use to conduct their research – to analyse data, to search databases, to run simulations and to record their scientific progress".

One interviewee described the role of software by analogy with "my view of an old fashioned chemistry laboratory, where somebody would go in and build a complicated array of glass tubes with interestingly coloured liquids and smoke coming out and that was their experiment. Software is the same: somebody constructs bits of code and links it together, and it may or may not produce some results, which may or may not be correct".

Scientific software is often developed by individual researchers who, with few exceptions, have had no training in software engineering, and this can give rise to problems.

This is borne out by this investigation, and indeed some of the most interesting developments are taking place in areas where computer scientists are working with the life science communities.

Good scientific practice is based upon the premise that results of experiments are can be replicated. As Lehrer [16]writes in the *New Yorker* "Replicability is how the community enforces itself. It's a safeguard for the creep of subjectivity. Most of the time, scientists know what results they want, and that can influence the results they get. The premise of replicability is that the scientific community can correct for these flaws. "

If the software tools used on (publishable) results are not made available it can follow that the results cannot be replicated.

At a slight tangent it is worth mentioning that scientific fraud is not unknown. Two interviewees raised the issue of cases of 'data fraud' in their own subject fields: in both cases results had been published and the authors had garnered the appropriate recognition. Eventually when no other research groups could produce these results, it was discovered that the datasets used were fraudulent. However it took years to uncover the deception. An editorial in *Nature*[17] in 2010 drew attention to scientific misconduct, citing a study where 2% of those interviewed had admitted that they had falsified, manipulated or modified data , but a staggering 14% had noticed this behaviour in colleagues.

An additional safeguard would met by the depositing of software code alongside data at the time of publication: however very few publishers currently require this.

Interviews conducted in the course of this project found that, without exception, all researchers believed there was a problem that some of the more intangible parts of the data-cycle had not been curated and thus was not accessible to anyone else. These may, or may not have benefits within communities or across domains.

As one physicist reported : "It was very obvious that there were PhD students, sitting in little cupboards tapping away at their terminals reinventing the wheel, reinventing little fitting procedures and data analysis procedures which were exactly what we had been doing at CERN 20 years before".

This was not a problem at the level of large collaborations (where the infrastructure is in place, and data and software preservation is largely integrated) but at the level of smaller research groups. This was equally true across all subject domains.

The code problem for smaller groups was endemic; as explained by one High Energy Particle Physicist (who works on the CERN Atlas experiment): "such as take place at the end of my corridor", where "a run of data will be analysed on a graduate student's PC: they will quite often will not like the result, and the code will be rewritten constantly….and it is difficult for a student to understand to what, if any (code) they should keep".

As a health warning it should be noted that, as with data, much software code should perhaps properly be disregarded: one interviewee making the point that often"someone coming to the data may in most cases be better off starting their software from scratch".

The point is that currently too much is destroyed or left to languish on a post- doc's computer.

From a policy perspective, the funding councils are increasingly aware of these issues: there is recognition that data itself is often not well managed[18] and the introduction of mandatory data management plans is attempting to rectify this in a 'top down' approach. As Neylon[19] states, the drivers behind this are political - the top down view from government that publicly-funded research needs to gain from the benefits they see in data-rich commerce.

A complementary or indeed alternative riposte to the issues of inadequate data management infrastructure (including computer code) can be seen in the 'bottom up' approach of scientists. This is particularly the case for the genomics community.

Comment [JN12]: Am slightly puzzled by the logic here. Has it been transcribed correctly?

Comment [y13]: It was oddly but have amended it to get the sense of it!

Comment [JN14]: Is this a quote? If so, where does it begin?

In this they are actually engaging with the problems, by attempting to ensure that the code they create or use, is more discoverable (for example by using ontologies or workflow management tools). The latter include researcher-led tools such as Lark ('simple framework for describing experiments, hypotheses, materials and methods relating to research, supporting everything from laboratory protocols to computational workflows') which is currently under development[20].

These organic approaches, percolating upwards from the research groups, provide an opportunity for librarians and other research support services to be involved different research support models at an institutional level.

An active role for libraries?

In 2003 Hey and Trefehen[21] (writing from an e-science perspective on implications for the library community) considered whether new types of libraries for scientific data should be created with the same sort of management services as conventional digital libraries.

They suggest that the increase in data due to new scientific techniques will need collaboration between scientists and computer scientists to "analyse, federate and mine this data", and that to "organise, curate and preserve this data" will require collaboration between researchers and libraries.

Arguably the role of information professional should be moved forward in the research process to that of researcher support in relation to 'live data ' and the techniques deployed on this to gain experimental results.

This is not to undermine the role of the librarian as curator: however the current research demonstrates that there is a role in advising on what data outputs an individual researcher or group should keep, and to provide assistance in doing so, and the requisite tasks to aid discoverability.

Other issues include researchers making a value judgement that a piece of software could potentially be of use to others and whether they should share it. Also relevant are the working practice of the researcher, their domain's culture and their institutional polices and whether these facilitate good data management practices.

In terms of facilitation there is clearly a role for information staff and other research support services. As one librarian interviewed said, "we (librarians) need to understand our researchers need, habits and workflows. Information management must take a role to help researchers to find the right tools for their needs".

More importantly, *all* of the researchers interviewed (and indeed all said it was the first time that they had had a conversation on this subject with any one resembling a librarian) believed that there was an active role for a library service in this respect.

A 2008 RIN study draws attention to the gap between the specialist roles of informaticians, statisticians, modellers and curators, and the information skills of life sciences researchers[22]. It concluded that engagement with information professionals could add to the efficiency and effectiveness of research in the life sciences. An equivalent project has recently been conducted with regard to the physical sciences (and which the author contributed to) and is

due to be published in November 2011. The JISC funded *Incremental* project, which was library based, aimed to identify current practices in managing digital research data, and to assist researchers in managing their data in the future.

It is also heartening to know that we (Cambridge University Library) are not the only library service attempting to come to terms with the changing research infrastructure and to determine what our role should be in assisting researchers manage their data outputs.

Comment [JN15]: "We" being Cambridge??

Research library staff from CERN are also dealing with these issues. One analysis from the CERN research defines the role of researchers is "as data providers (data production), data documentation, data submission (preparation of publishable datasets) and data quality assurance, and as data users their remit is the correct citation of the dataset."

The role of the information professional is the management of the research material remix, assisting in facilitating workflows, integration of new features (for example citations) preservation and cataloguing (making the data discoverable) and to assist in the submission of articles for publication[23].

During the course of the current research several scientists expressed their hope that alongside these services listed above information and other research support staff (within the Cambridge context this refers to the e-science centre) should play a more proactive role in training researchers in data management and workflows, use of version control software) and provide a service to curate software from research groups and to help make this discoverable.

Methodology

Comment [JN16]: I wondered whether this section on methodology should come earlier in the document?

Comment [y17]: The only other place would be before the background ?

The research aimed to examine how research data is currently handled by research groups in large scientific collaborations, and to question at what stage (if any) input from information professionals would be of use.

Two broad subject domains were initially chosen for investigation: physics and bioinformatics - candidates due to the large amount of data handled by these subject areas and to the fact that researchers have to use computer code to manipulate the data products. However this remit was widened following expert advice and several researchers undertaking multi-disciplinary work in the biological sciences were also interviewed.

In all cases it is fair to say that the selection of academic researchers chosen for interview was biased (those who were approached were known to have an interest in data management or publication or were undertaking multi-disciplinary work). These included academic researchers working in the areas of bioinformatics, biology and physics and those working in multi-disciplinary teams. These individuals were approached on the basis that they already had an acknowledged interest in data sharing, or were known to be undertaking inter-disciplinary research with collaborators from another field.[24]

This approach was purely pragmatic: time was limited and it was felt that researchers who had no real interest in the subject would not be willing to give up quite a considerable amount of time to be interviewed. The average interview length was an hour.

A series of interviews took place, the majority in person, the rest by phone, with academic researchers being asked a broadly similar set of questions.

These covered:

- The generation of redacted data (what processing takes place)
- How/if the techniques used on this 'live data' are recorded (if they are)
- What happens to bespoke computer code at the project and individual researcher level?
- Are the techniques / code made available to researchers in other organisations?
- Do the researchers see any potential re-use of the techniques they employ for other fields?
- How would they go about researching other data techniques that may impact on their research?

In total 30 interviews were conducted.

Other stakeholders - including librarians, policy makers, publishers and research scientists with a professional interest in open science - were also interviewed.

It should be noted that some of the author's initial assumptions were naïve and the interdisciplinary nature of this research (and the relatively short time scale of the project)

means that many of the results are merely touching upon the surface of what could be achieved with a longer-scale study.

It does however highlight the significance of cultural differences between disciplines whilst demonstrating that many of the potential problems are actually shared by all fields.

Interviews are grouped by discipline as these broadly share domain characteristics and practices.

Life Sciences

Genomics

A series of interviews were carried out with researchers, support staff and policy advisors working at the Wellcome Trust Sanger Institute (WTSI), a charitable-funded genomics research centre, and the European Bioinformatics Institute, part of the European Molecular Biology Laboratory (EBI-EMBL), an academic research institute located on the Wellcome Trust Genome Campus.

Genomics research generates data on a massive scale. Brooksbank[25] *et al* in describing EBI's data resources state that the genomic era has heralded a social change for the life sciences: the scale of genome sequencing meaning that biological experiments are now generating data at rates comparable to particle physics or astronomy. The WTSI has one of the largest sequencing facilities in the world. Facilitated by the next-generation sequencing platforms in use at the WTSI, the facility is now producing over 1 terabase (1000 billion bases) of raw sequence output every week and this is expected to treble within the next year.

A single DNA sequencer can now generate in one day what it took 10 years for the Human Genome Project to collect. The introduction of "next-generation" machines, faster sequencers that "spit out data more cheaply has meant the machines generate such short stretches of sequence—typically just 50 to 120 bases—that far more sequencing is required to assemble those fragments into a cohesive genome, which in turn greatly ups the computer memory and processing required. It once was enough to sequence a genome 10 times over to put together an accurate genome; now it takes 40 or more passes[26]".

This led to some working in the field to question whether the torrent of DNA data and the need to analyse this "will swamp our storage systems and crush our computer clusters".

Although highly unlikely, the costs of storage are dropping more slowly than the costs of generating sequence data; "there will", said one interviewee, "come a point when we will have to spend an exponential amount on data storage[27]. This means that questions will have to be asked about what data is stored (and whether the original raw data will be discarded).

It is now arguably more cost effective to generate sequence data as needed, than store the raw data. Such questions are outside the remit of the current research but it is worth bringing them to the reader's attention to highlight awareness of the issues being faced at an institutional level.

In the field of genomics the ethos of data-sharing practice is well established, with the field being "regarded as a leader in the development of the infrastructure, resources and policies that promote data sharing"[28].

The extensive RIN 2008 study '*To share or not share: research data outputs*', summarised the position of genomics in relation to data sharing as follows:

Comment [y18]: Starts here

Comment [JN19]: Where does this quote begin?

Comment [JN20]: Have you referenced this before?

Comment [y21]: Yes at the start

Culture of data sharing: high

Infrastructure related barriers to publishing data: Low

Effect of policy initiatives to encourage data publishing: High

Overall propensity to publish datasets (with appropriate metadata and contextual information): High

The data-sharing policies of funders in this field are arguably one of the major drivers behind this and there has been direct investment given to the infrastructure needed to support this. In this respect genomics differs significantly from the other biological sciences.

"Open access to all data is believed to accelerate advances in science, by making data freely available to all while ensuring that the expedient use of existing resources that have been funded by the public purse". In genomics this principle was outlined in the Bermuda Principle in 1996, and followed by the Fort Lauderdale Agreement in 2003[29].

This is reflected at the institutional level. The Wellcome Trust expects that, as an absolute minimum, researchers should make relevant data available to others on publication of their research. The Wellcome Trust supports unrestricted access to the published output of research. Specifically the Trust requires that electronic copies of any research papers to be made available through UK PubMed Central as soon as possible or at the latest within 6 months of the journal publisher's official date of final publication.[30]

The WTSI library assists in ensuring compliance with open access by submitting research papers to the relevant repository. The leading journal in this field (*Bioinformatics*) reports discoveries using computational methods and has a distinct section detailing these computer applications.

Comment [JN22]: Which library?

As Kaye *et al*[31] in the past data sharing has primarily taken place with known colleagues and been based on "mutual respect, trust and a common interest". With the current funder, data sharing policies the question has become for researchers how to share data. However even in a field like genomics, the situation can be ambiguous, as researchers in this area often interact with other scientists whose practices are based on data confidentiality.

Against this background it would seem that data sharing between those working in genomics is in hand and that sharing of software code would be an obvious next step.

Yes. However as Goble and De Rouretate, "Scientists to be successful must be fundamentally selfish" and illustrate this with a quote from Mike Ashburner, a Cambridge geneticist, which captures this perfectly. "Scientists would rather share their toothbrush than share their data".

Researchers, support staff and policy advisors interviewed at WTSI were aware of good data sharing practices both at the institutional and national level and to implement policy there is a permanent high level standing committee on data sharing.

An institute-wide data sharing policy has recently been developed and implemented. The latest version of this is available on the institutions website and as institute policy it has the same effect as a mandate[32].

Issues considered in drawing up and implementing such a policy included cost factors and the anticipated benefits from making data available. These benefits must outweigh the costs associated with long-term archiving and the effort involved in preparing data for submission.

Questions considered in the creation of this policy included whether summary data (as opposed to raw data) be stored, as these decisions necessarily affect what data to archive. Protocols for 'managed access' to data were also considered. The authors conclude that for the implementation of a data sharing policy to be effective, it must be carried out in systematic and comprehensive way; "it is easy for data sharing to be seen as a burden ….It is essential that the entire scientific community if researchers and funders is satisfied of the overall benefit of sharing data to science".

Within this domain it is apparent (and reinforced by interviews in the course of this research) that the key drivers to data sharing are the funders' policies and high the level of institutional support available. There is very little resistance to the idea of data sharing, for in this community "it is part of the ethos". "Sharing data makes sense, sharing software makes sense".

Annotated, open databases of genome sequencing are provided by the WTSI on its website[33], as is open access software developed in the course of research. Researchers are given guidelines as to which database they should submit their data to: the databases are generic by datatype both to aid to discoverability of data (and because this arrangement is more cost effective).

Comment [JN23]: its?

It is acknowledged that these databases and software listings are used widely by the genomics community but that the software is not likely to be discoverable (or indeed used) by those working in other domains.

From a research support perspective it is apparent that genomics has a developed, mature suite of repositories to store and curate data and this should, as one interviewee suggested, alleviate pressures on scientists in relation to data sharing policies (software being developed at the institutional level being included in this).

Long term curation of most datasets takes place at EBI-EMBL. This reflects the funders' acknowledgement that there was a need for such a facility - "a bit like a library for our datasets". Organisationally there are multiple data types and different databases for each type, and they need to be curated differently.

It was held to be essential in this field that the funders recognise the value of data management and fund such facilities. This is in stark contrast to the problems faced by those working in the biological sciences as described below.

From a policy perspective an attempt is being made to take the same approach to software code as the datasets : any piece of software that is useful is not submitted just to EBI but is made available from the WTSI website to facilitate the discovery of resources and software.

Facilitation of data and software is a concern and organisationally WTSI are trying to come up with better listings.

It is clear that although the priority of the researchers has to be the core science: there is an embedded cultural norm that data sharing, and to a lesser degree software sharing, must be facilitated.

One interviewee felt that data sharing at WTSI was exemplary due to the infrastructure in place: processed data was captured at the point at which it was created, enabling the application of meta-data. He stated that often the "technical structures were easy but the value systems around them were often missing".

Several researchers have introduced highly innovative ways of making data more accessible to both other researchers and interested readers.

One extremely interesting example is that all data collected in one database (Rfam[34]) a collection of RNA sequence families of structural RNAs including non-coding RNA genes, is made available via a Wiki in accordance with the principles of the WikiProject Computational Biology, which aims to organise and improve content in the area of Bioinformatics. Data annotations are exported into from the Rfam database into Wikipedia, and these Wikipedia databases are then reimported on a nightly basis into RFam, which essentially enables researchers for the first time to be able to directly edit the content of one the major RNA databases.

This initiative (described in detail in the RNA *Wikiproject: Community annotation of RNA families*[35]) reflects the researcher's understanding that Wikipedia has become one of the most important online reference sources with growing scientific content and that the route for many researchers to a topic is via Google and then into Wikipedia.

Accessibility or discoverability of existing software code was not seen to be a major barrier to progress: "Most of useful software is on genomics institutes are they are making it available through their websites if you know what to look for". Peer-to-peer sharing of software resources was not thought to be a problem by the researchers interviewed.

In research projects, code was written by the researchers or their colleagues. What was really important "is your data and your results, this is what you get your grant money for, the software is just a tool. Everybody is getting better at sharing data just cross domains, the software is not seen in the same way although I believe that bioinformaticians are better at sharing their software but we are a slightly weird breed".

Commonly code is written for a specific function: however as one researcher describes it, such code "tends to be quick and dirty". Iterations of code that worked for what it was written for was stored. Occasionally computer code was submitted to SourceForge, an open software repository. It was acknowledged that finding specific software code could be problematic and that the level of categorisation on SourceForge was not that helpful.

However the prevalent assumption is that "writing code for myself takes a few hours, making it available for others to use can take months". Moreover there is a judgement call as to the usefulness of computer code "To know what is going to be useful and to maintain it in a way that it is going to be useful is more difficult, more so than with the data"

The majority of researchers had no formal computer science training and several felt that this was reflected in their working practices, indeed one suggested that it would be useful to

take advice from computer scientists as it would be " interesting to see what they are doing in terms of software sharing as to them the data is interesting, but what is really is important (to them) is the code".

A clear weak point, according to all interviewees in this domain (and indeed all others), was software code written at the level of the PhD student or post-doctoral researcher - what one computer scientist working with the bioinformatics community calls the 'long tail scientist'. It was also noted in interviews that the area of training in good data practice was crucial and that this did not currently take place.

According to some senior computer scientists bioinformatics is "notorious for the reinvention of code". The leading proponent of this view is undoubtedly Goble[36]. Her keynote lecture, delivered in 2007, still resonates with the genomics community today, and arguably is true of all data intensive fields.

The seven sins 'deadly sins of bioinformatics' that she identified, with some illustrative examples, are listed below.

1. Parochialism and Insularity: encompassing reinvention, reinventing the wheel, rediscovering the same problems and the creation of yet another database, yet another ontology, another web2.0 site, another portal…

2. Exceptionalism: domain specific outcomes demand domain specific tools.

3. Autonomy or death! Researchers want the ability to change the interface/format whenever they like, despite the fact that "I have lots of users who depend on this".

4. Vanity: Pride and Narcissism or claiming to know everything about a subject domain and computers

5. Monolith Meglomania: or "my data is mine and your data is mine too" and the trouble with 'warehousing' data: data is deposited to 'rest in peace'

6. Scientific method Sloth: or "it's easier to think of a new name than use someone else's" and worrying about errors in experimental data but believing that derived data is also always true. This also covers producing irreproducible bioinformatics analyses: and the practical example is to try running experiments in Bioinformatics from 5 years ago.

7. Instant Gratification: "the quick and dirty fix" and encompassing 'hackery': "producing crap, non-reusable, software because only the biological results matter for publication X"

Goble appreciates that these factors are both technical and social, and that reuse of data and techniques is hard: as she puts it "a few months in the laboratory (or the computer) can save a few hours in the library (or on Google)", although she notes that computer scientists are guilty of some of these too.

This analysis, although domain specific, could and should also be applied to the other subject domains reviewed and the author intends to extend this analysis to the physical sciences.

In attempting to address these issues, Goble and her colleague have been using social networking and community collaboration techniques to build collaborative "e-Laboratories" for sharing data, models, methods and workflows. They focus in particular on the "long tail" scientist: that is post-docs and students scattered in research labs and universities.

These tools include myExperiment[37]a community repository and virtual research environment that supports the sharing and reuse of scientific workflows and other kinds of experiment plans and methods. It has over 4500 registered users and over 1000 deposited workflows from 19 different workflow systems. BioCatalogue[38] is a crowd-curated registry of web services for the life sciences with over 1700 service entries. SEEK[39]is a private community collaboration and asset sharing platform for Systems Biology models, data and protocols serving 120 research institutions throughout Europe. MethodBox[40] is a collaboration environment for sharing variable sets and statistical methods for analysis across social science survey data.

Comment [JN24]: Something missing before this footnote

Based on an assumption of “scientific naughtiness” Goble questions whether one should try to deal with it or expose it. “Transparency and accurate collection and reporting is vital, alongside provenance and this should help put an end to black box science”. It must, she argues, be presumed that other researchers use software and data and that they can add value to it (as the Rfam database wiki clearly demonstrates).

However due to cultural factors within the bioinformatics community, these issues are to some degree being addressed. They are clear examples of good practice where the engagement of bioinformaticians with computer scientists is driving best practice.

For example, there was an awareness that discoverability of software was potentially problematic, both within and out with the bioscience community, and that making software discoverable outside the community is essential for cross-domain interaction. Is making software more discoverable an answer? And if so, how should it be done?

This tack has been followed by researchers at EBI who are involved as joint PI’s (with the School of Computer Science, University of Manchester) on the JISC SWOP Project. This project was very much a ‘seeding project’ running between February and July 2001.

It aimed to release an ontology of software that describes at least 200 pieces of software. Ontologies (a representation of knowledge within a domain) are used at EBI to describe the experimental variables in the data. Ontologies provide a mechanism for capturing a community’s view of a domain in a shareable form, that is, a form that is both accessible by human and computationally amenable and that provides a set of vocabulary terms that label concepts in that domain. The terms should have definitions and be placed within a structure a structure of relationships: the most important being parent and child[41].

In Bioinformatics, gene ontologies are used to represent the gene and gene product attributes across species and databases[42]. Working practice at EBI dictates that for the biological data that is held, ontologies are used to help annotate the data and to add semantic richness to help with querying and to assist with different views of the data (‘tree-browsing’) and integration across the data.

The SWOP project grew out of an idea that an ontology could be created to describe software and the different way people were interacting with data using that particular software. There was, the researchers felt, an acceptance within the community that researchers use different vocabularies to describe the software tools that they create, or indeed how they use software in their work and for what purposes, and that a mechanism for formalising this would be extremely useful.

Comment [JN25]: Do you mean 'acceptance'? I find the sentence hard to understand otherwise.

The SWOP project seeks to develop a vocabulary that will help describe software used by the curation and data preservation community based on the understanding that the description of software is crucial in areas of digital preservation, service integration, text mining, service discovery for users and in describing the provenance of curated data in areas such as bioinformatics, other life sciences, the physical and social sciences and many more.

Curation of the uses made of software was also a main driver behind the project. It should enable users to answer questions such as "this is the sort of data I have, tell me what can handle it", or "I am looking for this sort of algorithm, this is the kind of analysis I want to do", the software equivalent to "is this liver cancer, is this mouse"?

The researchers also suggest that such ontologies should also include cost of software and cover commercial software and licensing restrictions: on the basis that not all useful software is open source (although all software developed at WTSI and EBI is). The SWOP methodology is also scalable to different disciplines.

The participants in this project are keen not to repeat what one of them termed "the crime of a silo of ontologies". This is neatly demonstrated (as is indeed is the whole problem) by the fact that during the initial project planning/primary study stage they became aware of another project, also based at EBI, tackling many of the same issues. This project, EDAM (EMBRACE Data and Methods) is an ontology of general bioinformatics concepts, including topics and data types, formats, identifiers and operations. EDAM provides a controlled vocabulary for description in semantic terms of concepts strictly in domain of bioinformatics. General computer science or biological terms are (typically) not modelled. Where software has been modelled by EDAM that is now also used by SWOP.

Comment [JN26]: Eh? Keen to repeat a crime????

Comment [y27]: Oops!!!

Despite the problems inherent in a culture where sharing is officially the norm, but perhaps not always at the level of the individual researcher in relation to the computer code outputs of their research, the infrastructure of the WTSI and EBI-EMBL do make it easier for the data and code to be shared.

What is most interesting is that there is awareness of the issues and clear evidence of 'bottom up practices' such as SWOP and the Rfam database, where researchers are tackling issues relating to research data and outputs. Many of these practices should be disseminated to other domains, as should the analysis of the 'deadly sins'.

Biological Sciences

Researchers within the School of Biological Sciences were also interviewed. These individuals had been recommended to the author by the Cambridge e-science centre as

their work involved the creation of large datasets either by the use of NMR spectroscopy, X-ray crystallography or time-lapse microscopy. All of these had had some level of involvement in projects the Cambridge e-science centre.

These interviews served to highlight the cultural differences between different subject domains and provide a complete contrast to Genomics.

In comparison, these researchers were not supported by large-scale infrastructure and were often trying to find solutions to problems which would have been resolved by access to the database and curation facilities available at WTSI and EBI-EMBL.

RIN's 2008 study, *To share or not share: research data outputs*, summarised the position of systems biology in relation to data sharing as follows:

Culture of data sharing: medium

Infrastructure related barriers to publishing data: medium

Effect of policy initiatives to encourage data publishing: high

Overall propensity to publish datasets (with appropriate metadata and contextual information): medium

The major funding council in this area (the BBSRC) provides that publications must be deposited at the earliest possible opportunity and data must be made no later than the release of main findings through publication, or three years as a general guide. Researchers must submit a 'Data Sharing Plan' as part of their proposal. Specific scientific areas have established best practice for release of data as noted in the BBSRC Data Sharing Policy[43] .

However although funding can be claimed for infrastructure support, existing projects have not yet seen the benefit of this.

In all cases, and by direct contrast with genomics, curation of data was not straightforward: one interviewee called a it a "huge problem" to keep track of all the data that it is generated and the storage of this data and the associated metadata.

To alleviate this one researcher was actively trying to create a lab management information system, with each group having a centralised data store within their lab with details of all the different projects. It was believed that better curation of this data would enable better data sharing practices.

They were also actively considering having researchers output their results to Wikipedia in a manner similar to that undertaken by the Rfam database group at WTSI, on the basis that if you wish to make data available it should be in a form that people can easily read and annotate. Interestingly *Google* was cited as a tool for the democratisation of scientific data (although this assumes the data has been made available in the first place!).

The data sharing culture was, unlike genomics, far less developed: "In structural biology a great deal of intermediate data is never deposited, We have fundamental problems convincing people that they should share data. Personally I think the only way it can be overcome is by the funding bodies and the journals insisting on it….we also would like a

consensus in the community; it is happening gradually, but it is slow and for it to work you need to come up with a way of connecting software to data, But there is a long way to go to convince scientists that they should do so".

Some considerable envy was apparent at the set up at WTSI as expressed by statements like "In genomics people have got their act together". It was felt that the pervasive agreement/ mandates determining the public availability of data were the main incentives and that "once you get away from that, the ethos of sharing is not so strong".

Although there was general agreement that it would be beneficial for the community for researchers to deposit or make available software code, many were hesitant to do so, for as one interviewee reported that researchers in his group were "scared" of making code available due to the fear of errors being discovered in that code.

Comment [JN28]: Gosh! That's a very revealing comment, given the scientific ethos of being open about everything, including one's apparatus.

One researcher who was interviewed was extremely keen to interact with experts in other domains relating to the techniques that his group were using. He discussed at length his problem with sharing data and techniques and this case illustrates many of the barriers to data sharing.

In his field (using time-lapse microscopy) there are a number of significant problems with sharing data. The outputs of such research tended to be in different formats reflecting the different microscopes used, and it is therefore not necessarily straightforward to know what the format of the data is to share it. There are publicly-funded initiatives to try and cope with this sort of problem: the Open Microscopy Environment[44] (OME) is funded by the Wellcome Trust and NIH. This attempts to provide a framework for supporting data management for biological light microscopy. It is designed to interact with commercial software and all OME formats and software are free, and all OME source code is available.

However OME is not used by this particular research group: the problem from this group's perspective is that a great deal of time is spent creating movies and developing the code to do the requisite analysis. Doing this in someone else's developmental environment is not ideal because " you don't want to be reacting other changes, so we end up writing our own, just for the purposes of dealing with it".

This is not the sort of project (due to sensitivities surrounding the subject area) which can be put into the public domain. However they do collaborate with other researchers "as soon as techniques become known we pick up collaborators". The development of computer code is not viewed as a 'stand alone' activity but as part of a process that fits within a larger package.

However, code does not live in a vacuum! As they have found "If publish your code in the public domain, (when in reality it was created for your own working needs and requires data, often a particular kind of data) , it may be quite useless to other users. However when other's want to use it, they will most probably have queries and will expect support".

In their experience it has proved extremely difficult to get funding for the development of computer code (even those there are many in their community who desperately need such resources).

Comment [JN29]: Is this passage a quote?

This group is working with the Cambridge e-science centre in an attempt to create a multi-disciplinary image-processing centre, starting with microscopy. This would facilitate individuals from different departments with the required skills (relating to storage, annotating, archiving and image analysis) to come together, with attendant economies of scale, This is driven by the group's need for computing resources for storing images, and the supporting infrastructure, but also an awareness that others within the University must have skills in this area .

However it is "difficult to make bridges" outside one's own domain and this experience was common to all the biological scientists interviewed. As one said "It can be difficult to find someone to engage with your problem you really need biologists and numerical people coming together and talking then the technical stuff becomes too abstract and technical and the biologists don't have the right grasp of the right physics questions to ask."

However there are good illustrative examples of novel research occurring by serendipity: "An awful lot of actual research progress is entirely serendipitous: it relies on people meeting each other"

An interesting example of such a serendipitous collaboration is the application of the techniques of molecular evolutionary biology to the analysis of a range of texts from the Bible to Chaucer's Canterbury Tales[45]. This is a truly multidisciplinary partnership involving manuscript scholars from around the world.

Comment [JN30]: What a fascinating example.

This origins of this project were a conversation between a manuscript specialist and the lead PI who realised that there were parallels between how mutations accumulate in DNA sequences as they evolve and how changed were incorporated into manuscripts when they were copied by scribes in the days before printing.

An issue that this draws attention to is funding: if one is funded to do research in a particular area, and you find yourself applying your techniques in another field, is this actually advantageous or otherwise? Publications outside an author's core field will not usually count towards their department's REF submission. Problems with funding are pervasive (and critical!). Some funding was forthcoming for the microscopy project from the EPSRC for applying physics techniques to biological problems.

Comment [JN31]: On what criteria? RAE and REF?

Increasingly, academics are judged by their publication record: at the most basic level publications in other than a scientist's main funded area of activity may not be counted, and even if they are, they are not 'core' to the funding received. Increasingly the Research Councils are 'contracting back' their research funding to core activity, with multidisciplinary work far less likely to be funded (as will be seen later in examples from the physical sciences ,including with the PathGrid project).

Comment [JN32]: What does this mean?

Moreover scientists are not rewarded for the development of computer code, in an era where citations equate to the value of research, this can mean that for many researchers the incentives are not there. In the biological sciences techniques used are commonly cited but this does not happen with computer code. This is clearly inequitable, and puts the current reward system offered to academic researchers under scrutiny: as one author states, "if someone else came into my lab and used my bench space that would be a collaboration" where if they used code that is developed by a research group that will often not be acknowledged.

Academic publishers are aware of these issues: increasingly journals require data to be submitted with experimental results; however what is important is the citations that authors receive when their article is referenced. However the deposit of data does not assist in this academic metric -- which can help determine researcher's careers. *Nature Protocols* publishes the protocols being used to answer outstanding biological and biomedical science research questions, including methods grounded in physics and chemistry that have a practical application to the study of biological problems. These are peer reviewed, fully edited and styled prior to publication.

The issue of incentives is key: how does one reward researchers for doing novel work in a multi-disciplinary arena, or at least ensure that they are not penalised for doing so? A heated debate took place on the question at *Science Online (London) 2011*, where Cameron Neylon expressed these frustrations: "Yes we needed to talk about the challenges and surface the usual problems, non-traditional research outputs and online outputs in particular don't get the kind of credit that papers do, institutions struggle to give credit for work that doesn't fit in a pigeonhole, funders seem to reward only the conventional and traditional, and people outside the ivory tower struggle to get either recognition or funding. These are known challenges, the question is how to tackle them"

Comment [JN33]: Something missing here

As yet, on this issue the jury is still out.

[1] The Data Big Bang and the Expanding Digital Universe: High-Dimensional, Complex and Massive Data Sets in an Inflationary Epoch
Authors: Meyer Z. Pesenson (1), Isaac Z. Pesenson (2), Bruce McCollum (1) ((1) California Institute of Technology, (2) Temple University) (Submitted on 3 Mar 2010)

arXiv.org > astro-ph > arXiv:1003.0879 last accessed 25/10/2011

[2] Is There an Astronomer in the House? Sarah Reed
Science 11 February 2011: 696-697.[DOI:10.1126/science.331.6018.696]

[3]The World's Technological Capacity to Store, Compute and Process Information
DOI: 10.1126/science.1200970
Science **332**, 60 (2011);
Martin Hilbert*, et al.*

[4] http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204
Last accessed 25/10/11

[5] http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-WP3-DEL-0002-1_0_public_final.pdf
Last accessed 4/11/2011

[6] YN has interviewed and subsequently been in correspondence with one of the authors of this report in advance of publication.

[7] http://royalsociety.org/policy/projects/science-public-enterprise/
Last accessed 17/11/11

[8] Science as a public enterprise: the case for open data

Geoffrey Boulton, Michael Rawlins, Patrick Vallance, Mark Walport
The Lancet, Volume 377, Issue 9778, Pages 1633 - 1635, 14 May 2011

[9] What Researchers Want, Feijen M, 22.2.2011 Surf Foundation
http://www.surf foundation.nl/nl/publicaties/Documents/What_researchers_want.pdf

[10] Managing Research Data –Gravitational Waves'
DRAFT Final Report, Norman Gray, Tobia Carozzi and Graham Woan

University of Glasgow https://dcc.ligo.org/public/0021/P1000188/006/report.pdf

[11] Sherpa Juliet Research funders archiving mandates and guidelineshttp://www.sherpa.ac.uk/juliet/

[12]Memorandum submitted by Professor Darrel Ince (CRU 34)
http://www.publications.parliament.uk/pa/cm200910/cmselect/cmsctech/memo/climatedata/uc3402.htm-
Last accessed 1/11/11

If you're going to do good science, release the computer code too
Darrel Ince
guardian.co.uk, Friday 5 February 2010 15.42 GMT
Last accessed 1/11/11

[13] With thanks to Matt Wood

[14] Data in Motion: a new paradigm in Research Data Lifecycle Nicholas F. Tsinoremas, Joel Zysman, Christopher Mader and Jay Blaire ...

www.columbia.edu/~rb2568/.../Tsinoremas_UMiami_RDLM2011.pd...File Format

Last accessed 12/09/2011

[15] Six Principles of Software Design to Empower Scientists Goble
David De Roure, Carole Goble
IEEE Software (January 2007) Key: citeulike:2801066

[16] The truth wears off,is there something wrong with the scientific method, Jonah Lehrer (Dec 13,2010)
http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer#ixzz1drc9yeNK

[17] Combating scientific misconduct
Nature Cell Biology 13, 1 (2011) doi:10.1038/ncb0111-1
Published online 21 December 2010

[18] Interview with JISC director of a major funding stream

[19] Interiew with Cameron Neylon, STFC and commentator on open science issues
http://cameronneylon.net/

[20] http://www.larksong.org/introduction/

[21] Tony Hey, Jessie Hey, (2006) "e-Science and its implications for the library community", Library Hi Tech, Vol. 24 Iss: 4, pp.515 - 528

22 Case studies in the life sciences: how researchers use and manage information resources RIN

http://www.rin.ac.uk/our-work/sing-and-accessing-information-resources/patterns-information-use-and- exchange-case-studie

[23] Research data "publishing"-models, roles and responsibilities
Expert Conference on openaccess and open data, Cologne 13-14 December 2010
Sunje Dallneier-Tiessen, CERN Geneva

[24] Advice on suitable candidates for interview was taken from several academics who were working in multi-disciplinary areas of science and the Cambridge-Science Centre.

[25] Brooksbank C., Cameron G., Thornton J. (2010)

The European Bioinformatics Institute's data resources. Nucleic Acids Research 38: D17-D25.

[26] Will Computers Crash Genomics? Human Genome 10th Anniversary

Elizabeth Pennisis *Science 11 February 2011:*
Vol. 331 no. 6018 pp. 666-668

[27] Tim Hubbard quoting Ewan Birney at Science Online 2011.

[28] Nature reviews genetics, May 2010, Vol 5

[29] Data Sharing in Genomics – Re-shaping Scientific Practice

Nat Rev Genet 2009 May: 10 (5) 331-335

Jane Kaye, Catherine Heeney, Naomi Hawkins, Jantina de Vries, and Paula Boddington

[30] Sherpa/Juliet
http://www.sherpa.ac.uk/juliet/

[31] Nature Reviews Genetics 10, 331-335 (May 2009) | doi:10.1038/nrg2573
Data sharing in genomics - re-shaping scientific practice
Jane Kaye1, Catherine Heeney1, Naomi Hawkins1, Jantina de Vries1 & Paula Boddington1

[32] Developing and implementing an institute-wide data sharing policy
Stephanie OM Dyke, Tim JP Hubbard
Genome Medicine 2011, 3:60 (28 September 2011)
http://genomemedicine.com/content/3/9/60

[34] http://rfam.sanger.ac.uk/
Last accessed 21/10/11

[34] Rfam: Wikipedia, clans and the "decimal" release

Nucl. Acids Res. (2010) doi: 10.1093/nar/gkq1129

[36] Goble's Wikipedia page
http://en.wikipedia.org/wiki/Carole_Goble
Last accessed 21/10/11

[37] (http://www.myexperiment.org

[38] (http://www.biocatalogue.org)
Last accessed 21/10/11

[39] SEEK (http://www.sysmo-db.org)
Last accessed 21/10/11

[40] (http://www.methodbox.org)
Last accessed 21/10/11

[41] STEVENS RD; LORD P; BRASS A; GOBLE C. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics. 2003 July; 19(10): 1275-1283.

[42] http://geneontology.org/
Last accessed 21/10/11

[43] BBSRC Data Sharing Policy
Last accessed 21/10/11

[44] http://www.openmicroscopy.org/site
Last accessed 21/10/11

[45] See for example http://www.newton.ac.uk/programmes/PLG/Abstract3/howe.html
Or listen to
Do manuscripts drift like DNA (the Naked Scientists)
http://www.thenakedscientists.com/HTML/content/interviews/interview/1090/