Create, curate, re-use

the expanding life course of digital research data

Chris Rusbridge
Digital Curation Centre

Abstract

Scientific communication used to be based on the article or the monograph. Now datasets and databases are becoming as important in some cases. Aside from their value in communication, data are also the raw stuff of the scientific record, and the basis for verifiability. So scientists need to curate the data they create, and make them available for re-use. What are the implications and effects of these changes, and what should scientists and scholars be doing about them?

Curation

Curation is not a new term, being well established particularly in art and museum practice. However, it is relatively new in relation to data. We are now generally well aware that there are issues relating to the long term preservation of digital data (known as digital preservation), but digital curation is more than this: maintaining and adding value to a trustworthy body of digital information over the life-cycle of scholarly and scientific materials, for current and future use. Implicit in this is that data are thoughtfully created, carefully managed and curated, and re-used in a disciplined way, where and when appropriate. Also implicit is that curation is a whole life process, with potentially evolving digital objects.

Curation is clearly domain-dependent, with significant issues relating to size, numbers of objects, complexity of objects, interventions needed, ethical and legal implications, policies, practices, standards and incentives.

The Digital Curation Centre (see http://www.dcc.ac.uk) takes a broad view of digital curation. Whilst not exclusively data-oriented, we predominantly focus on data resources for science and scholarship. We are concerned with:

- The sustainability of the resource.
- The creation or appraisal, selection, acquisition and ingest of the resource,
- Growth, development of and changes to the resource,
- Making the resource available (“publishing” it),
- Access management and other controls on the resource, and the ethical and legal basis of these controls,
- The ability to use, combine, re-combine, inter-operate, process, annotate, discuss and review the resource through time (some of which processes will in turn contribute to the development of the resource),
- Linkage, context and metadata relating to the resource,
- Maintaining authenticity, integrity, provenance and computational lineage information relating to the resource,

- Maintaining the meaning of the resource despite technology change and concept drift in the outside world,
- Preserving the resource, including preserving access to past states of a changing resource,
- De-selection and deliberate and/or accidental destruction of the resource.
- All of this, over potentially extended time periods, although timescales could also be comparatively short or medium term;
- Recognising the impacts of finite budgets and potential future policy changes, and
- Paying attention to the education, training and development of the people to support this.

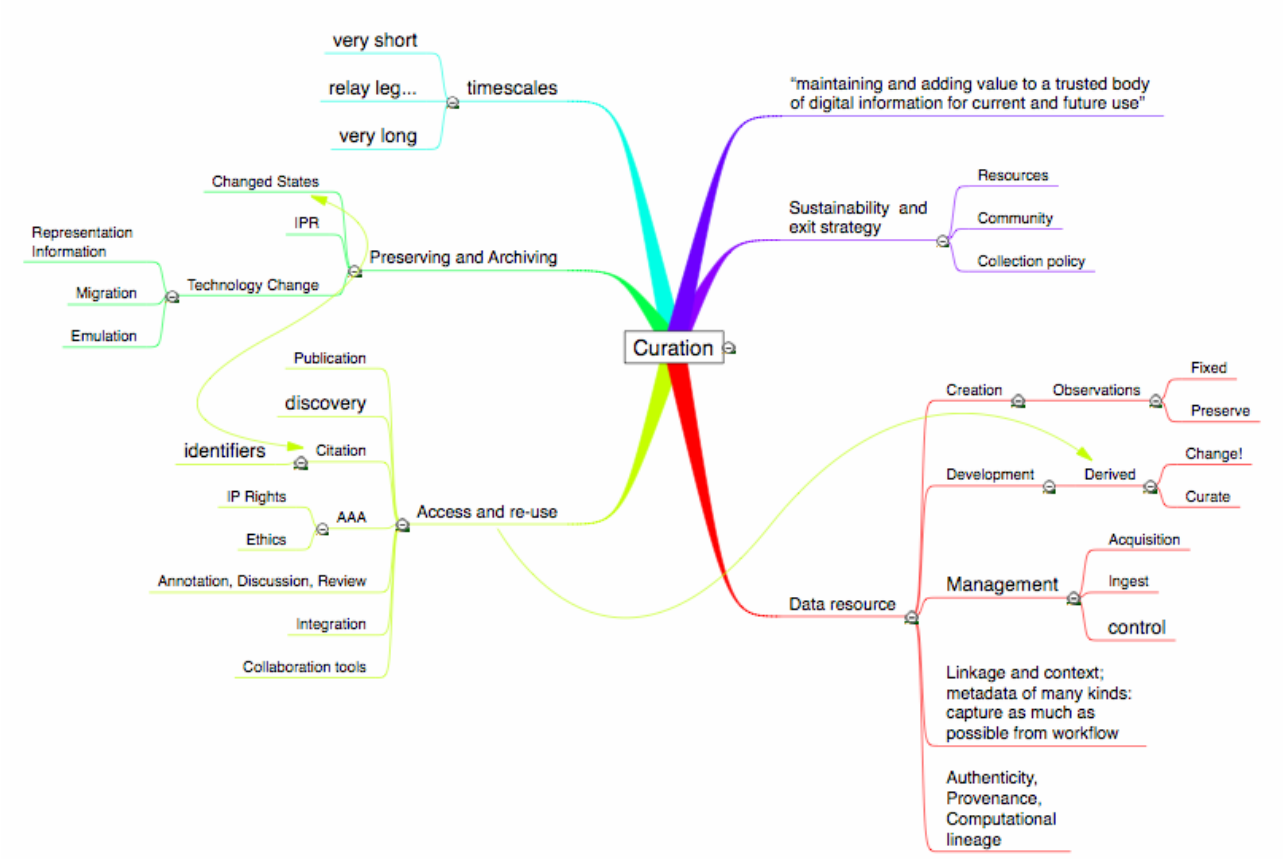

Figure 1: Mind map of good practice in curation

Good curation brings good science

Science increasingly depends on large quantities of data, representing experimental observations and/or other forms of derived or captured data. Managing data of this kind requires discipline if the results are to be scientifically useful. In some shorter term projects, with stable staffing and good communication, the group's "common knowledge" and good sense are sufficient to manage their data well enough to produce sound scientific results. Too often, one or more of these does not apply, and the result is data which leaves even its own investigators scratching their heads: data shorn of their context, of the associated experimental conditions, in un-documented files, in convoluted spreadsheets whose authors have left…

Managing your data properly simply means keeping the necessary context information and associated documentation to make sure you and others can make use of your data when the need comes. Good curation means good science… and conversely, poor curation may easily link to sloppy science.

Some science impossible without curation

Some kinds of science are impossible without data from the past. Here are a couple of examples from correspondence. Note, these do not necessarily represent well-curated data.

- "A prediction of quantum chromo-dynamics (QCD) is that the strong coupling 'constant' (α_s) should be large (~1) at low energies and small (~0.1) at high energies. In the 1990's the superb data from LEP at high energies showed α_s is small. The lower energy data came from older experiments Tristan in Japan, PEP and PETRA in Hamburg, and Doris and SPEAR. Sigge (Siegfried) Bethke wanted to demonstrate the running of α_s, and was reduced to hunting round various laboratories for old data. He did an heroic job obtaining the data, finding equipment to read it, contacting people to find out about the data format (not much encapsulated metadata there) and eventually produced a beautiful plot shown clearly α_s 'running' in the expected way." (Private communication from Ken Peach, 2005, referencing (Bethke, 2000))
- "In the 12th century BC Shang dynasty Chinese astronomers inscribed eclipse observations on 'oracle bones' (animal bones and tortoise shells). About 3200 years later researchers used these records, together with one from 1302BC, to estimate that the total clock error that had accumulated was just over 7 hours, and from this derived a value for the viscosity of the Earth's mantle as it rebounds from the weight of the glaciers." (Private communication from David Rosenthal, referencing (Pang et al., 1995)).

Why data?

Why should you bother to curate your data carefully? Some of the reasons have already been suggested: your own project will benefit from good management of your data, particularly as its volume (in both size and numbers of observations, files, databases, interpretations, metadata, workflows) increases. Here are some more reasons.

Important part of scholarly record

Fundamentally, data are an (often un-recognised) important part of the "record of science". I mean this in several senses: as the "public good" resulting from investment in your science project, as the evidence on which your findings are based, and sometimes formally as a record in the legal sense (Buchhorn and McNamara, 2006).

Scientific reputations are built today on published articles, and the influence those articles exert, as measured by the prestige of the publishing channel and the citations the articles attract. Shotton (Shotton, 2006) has pointed out that scientific articles are exercises in rhetoric, designed to persuade the reader of an hypothesis, see also (de Waard and van Oostendorp, 2006). Within a basic framework of scientific integrity, data are brought in as

supporters in this rhetorical structure. There is no shame in selecting the most appropriate data to illustrate a point, and no space to include all the possible data that might support (or contradict) the analysis presented. There would be great advantage if data are routinely made accessible so that findings can be substantiated by independent observers.

In fact, major medical publishers have made registration of clinical trials a requirement before publication (ICMJE, 2006). Likewise the International Union of Crystallography (IUCr) requires deposit of crystal structures before publication of articles referring to them. In the latter case, IUCr has also acted as a "Community Proxy" (NSB, 2005) in establishing the CIF common standard form for crystal structures (IUCr, 2002), which they require all authors to use in depositing structures before their articles will be published. Publisher mandates such as these can be among the strongest motivators encouraging curation of data and deposit for re-use.

Similarly, funder mandates are increasing; many funders now require at least a data management plan, and several require deposit of data into publicly accessible data repositories, sometimes after am embargo period to allow the PI time to exploit them. The UK Economic and Social Data Service has reportedly found its update increasing after the ESRC mandate was changed from *deposit* to *offer for deposit*. Now that they get to choose if the dataset is worthy of retention (using a peer review system, see http://www.esds.ac.uk/aandp/create/depintro.asp), it is seen as a mark of esteem!

Opens up additional interpretation possibilities

The examples quoted above show how additional possibilities can arise to re-interpret data for other purposes. There are risks, but also substantial benefits in this.

Legal Compliance

Legal regimes vary from country to country; the Australian Partnership for Sustainable Research (APSR) survey report (Buchhorn and McNamara, 2006) mentions some of the legal compliance issues that may apply in Australia, where the definition of a record is apparently wider in scope than in the EU, with the consequence that archives and records legislation may apply to the records of science. Certain other compliance regimes are effectively exported internationally, like it or not, for example US FDA requirements.

Incentives?

One the problems with curation is that it is clearly regarded as that extra burden, the one just beyond what is currently possible, in the queue behind meeting the conference deadline and writing the grant application. An activity, it might appear, that is obviously worthy, but which costs time and effort, and does not pay back in academic currency (reputation-building articles). This paper attempts to show that it does pay back, but if data are to be more useful in the long term, as surely must be the case, we need to build more closely geared incentives.

Publishing in databases needs to be as well-recognised (in some cases) as publishing articles. Citation of data needs to become the norm, so that the contribution of on person's data to another person's research is clearly recognised. These attitudes have to be built into systems such as the Research Quality Framework, and also into internal promotional arrangements.

What kinds of data?

Experiments like those associated with the Large Hadron Collider at CERN, and even larger quantities likely to emerge from future experiments make it clear that even with storage at ever lower costs, we cannot keep everything. For many experiments with smaller quantities (up to a few Terabytes!), we can keep all the data, but it may not be appropriate to do so. Criteria for appraisal and selection then become an issue (see archival practice for examples, see (InterPARES, 2006) and (Ross, 2000).

Environmental Sciences have produced a number of systems to describe the "level" of data, see for example the British Atmospheric Data Centre description of UARS (Upper Atmosphere Research Satellite) data levels at http://badc.nerc.ac.uk/data/uars/levels.htm. Here level 0 is the raw output data streaming from telemetry and instrumentation, effectively at the level of voltage changes; it is devoid of context. Level 1 data has been converted to the physical properties being measured, but will still be in formats tied to the instrument. Level 2 is post-calibration, and would refer to entities such as calculated geophysical profiles. Level 3 would be gridded and interpolated, and at this level there might be no clear correspondence with any observations (but there should be a clear computational lineage or provenance path linking these steps). Bose and Frew (2005) report similar but slightly different level descriptions, from NASA.

One may need to keep more than level 3 data; level 2 might be sufficient after a time, but level 1 may be needed if there is doubt about calibration.

I have heard several suggestions from particle physicists that only higher level data should be kept, not just because of volume, but also because they feel that no-one outside their group would understand the combinations of instrumentation, calibrations and parameters required to re-do these analyses. So much for the verifiability of science?

Combined and crafted data

The biological sciences are now very well supported by at least 968 databases (110 more than last year (Galperin, 2007)) that combine data from many sources, and other (or sometimes the same) databases that are intensively hand-crafted by teams of skilled data curators.

Annotations on other people's data are becoming valuable currency (X asserts this is evidence of gene A while Y asserts gene B). Annotations can take many forms, but in bio-informatics databases, the proportion of manually curated records is decreasing, as the volume rises and more automatic (and inherently less reliable) techniques are used. Annotation is well-used in the bio-sciences, but is spreading to other disciplines (Bose, 2006)). Manual annotations can cause interoperability problems because of the lack of (or use

of different) controlled vocabularies, for example the lack of direct compatibility between the National Library of Medicine's Medical Subject Headings (NLM MeSH) and the Gene Ontology, GO (Kersey and Apweiler, 2006).

Descriptive (meta)data

Unlike text, many kinds of data are not self-descriptive. We all make efforts to give our data some metadata, whether this is in the file naming and folder structures we use, or (more rarely) by filling in Properties boxes, or by assigning keywords, writing abstracts, thinking up useful titles. With text we can often get away with it (and smart indexing utilities like the Mac Spotlight on local drives, and Google on the web help us overcome our shortcomings here). But with data we have to be systematic. For any substantial project, file names alone will not be enough; some supporting metadata infrastructure will be needed.

Provenance data form valuable parts of the context as well. Particularly in the case of manually curated data, we may need to ask "where did these data come from?", "what is the status of their source?" (Batterham et al., 2006). (Buneman et al., 2006) are looking for underlying technology improvements in database management to make this task easier.

In addition, in the case of data we may be specifically concerned to understand the algorithms and calibrations used to compute derived values, sometimes known as computational lineage (Bose and Frew, 2005).

How to curate it?

In a short article like this, it is impossible to give comprehensive advice on how to curate your data. However, here are a few suggestions on what to do and what to keep.

Build curation/re-usability into your workflow

One of the best pieces of advice (from the R4L project in the UK, see http://r4l.eprints.org/ (Coles et al., 2006)) is to make life easier for yourself, by building workflows to manage your data collection and processing pipelines. You will be more systematic as a result, and your data will be more reliable. You can also easily build provenance information and associated metadata into your workflows, and hence into whatever metadata catalogue or data structures you have built. Ironically, while your project is active, all the information anyone would ever need to re-use your data is all around you. Unfortunately, because "everyone knows", some is never written down until near or after the end of the project, when the post-doc has left for that great job in the US, and the PhD student has left for a merchant bank (she's a rocket scientist, after all), when the PI realises he hasn't a clue what some of those data files actually contain.

Whether you use formal workflows or not, make sure you capture everything you can while it's easily accessible; this might include some of the text and key parameters of the proposal to your funders, or items as apparently unconnected as health and safety plans and records (which may record who was doing what, where and when). You must of course keep and manage

your experimental parameters and calibrations your data file descriptions, database designs and schemas, tag libraries, questionnaires, etc etc…

Keep data, and the ability to process it

Of course you have to keep the data, preferably in standard data formats and file types, processed with standard programs. Open source has advantages over proprietary code in some (but not all) cases, as you are not forced to move forward onto new versions, and so should be able to recover data for longer. Home-crafted code is necessary quite often, but has risks (see Geoffrey Chang's retraction of several articles, due to a simple programming error that flipped columns in a table (Chang et al., 2006)). Make sure you keep the code, and that it is well-documented, commented and annotated. Preferably get someone else to maintain it before the author leaves. Don't trust this to your PhD student without good supervision. This one is important!!!

Make ownership and allowable uses clear

You will need to establish and often to document the issues relating to use of your data. In many cases involving human subjects, you will have to have your experiment cleared by your ethics committee. Your proposal to them and their response are critical records: keep them! If you make agreements about sensitive data with particular groups, these are promises on which future trust depends. Make them carefully and with what forethought you can, keep them, and make sure others will continue to respect them.

If you have partners in other institutions, you need to avoid disputes over data ownership. Make some sort of agreements and document them (at least in minutes or notes of meetings, if getting formal legal agreements presents too great an obstacle).

Make it citable

Relating to incentives above, make it clear how your data should be cited (follow standard formats and discipline practice as closely as possible), and cite both your own data and that of others. The best way to get credit in the academic world is to build a base of data citations. Many archives, eg those in the Social Sciences, specify how their data should be cited, but often at the dataset level. Peter Buneman (Buneman, 2006) is exploring how datasets should be cited at a finer level of granularity, and in the face of change (2006). Standards in this are often mildly contradictory and very little followed; the best is probably the NLM Internet Supplement (Patrias, 2001).

What re-use issues?

Not to be too alarmist, but re-using other people's data also brings problems. The first relates back to the issue of articles being rhetorical exercises. Once again stressing the integrity of good scientists, but data are collected for a purpose (data are not neutral with respect to the hypothesis being tested), and this can affect what is collected and how, and also how it is subsequently treated. Not all of this information will necessarily be clear from the documentation provided, so extreme care must be taken not to misinterpret

the data. In the cases quoted above, the scientists involved went to great lengths to take account of these issues.

Since data are not self-describing, it can be hard to find the data you need. In some cases, tools like Google will help (for example, (Murray-Rust et al., 2004) have reported good results with known item Google searches using an InChI, a globally unique chemical identifier). However most often Google will be useless for data. This is where storing good metadata is essential, but you must also know where to look. Once they are better established, following data citations from literature should be helpful; at the moment, Buchhorn (2006) implies that many scientists track down data by inference from following up articles.

Once you have found the data, it may prove difficult to understand them in the sort of detail required to analyse, and particularly to integrate with other data. Once you have understood the data, it may prove hard to use them; details like formatting, keys, defaults, truncation etc can get in the way. In these cases, well-documented data made available according to community agreed "standards" will be much easier to use.

Overall, it can be hard to know the risks and pitfalls. Nevertheless, for the right problem, using other people's data is essential, and can greatly extend your work.

Who does it?

Given this critical importance, how can we assure the continued curation of data? Whose job is it anyway?

There is some evidence that "big science" is comparatively safe. No-one gets a big science grant these days without a data management plan. This particularly applies to large international collaborations with shared instruments, as in astronomy, particle physics, oceanography, etc. However, James M. Caruthers, a professor of chemical engineering at Purdue University has claimed 'Small Science will produce 2-3 times more data than Big Science, but is much more at risk' (Carlson, 2006). The lone scholar or small group, working comparatively isolated is under great pressure to publish. Such a group, as noted in Buchhorn and McNamara (2006) will tend to regard data curation as a set of optional activities to completed once the pressure is off… and it never is! The data are often on individual or at best shared drives. They will often not even be adequately backed up. The individuals concerned are intimately involved in the scientific work; they know so much that they do not feel a need to write down: they know too much, and are too busy, to create good metadata or documentation. At best some time after the PI has moved attention on to a successor project, at worst when a staff member leaves and the accounts are deactivated and then deleted, these data will simply disappear; they have no tomorrow.

Perhaps not complete, here is a classification of data curators:

- Individuals, using their hard disks, or perhaps networked drives
- Departments or groups, whether using separate or shared drives
- Institutions, perhaps in the shape of their libraries

- Communities of institutions, either formal (as consortia), or informal (as in the case of the LOCKSS system, Lots Of Copies Keep Stuff Safe, a distributed service founded at Stanford University, see http://www.lockss.org/)
- Disciplines
- Publishers
- National services, perhaps national libraries or archives, or national data services, and/or
- Other 3rd party services.

In a book chapter to be published shortly, I argue (Rusbridge, in press) that institutional curation repository solutions have some fundamental sustainability advantages, but lack the necessary critical mass of domain science involvement in curation. Discipline curation services do exist at the network level, and have huge advantages for data curation in being able to direct domain expertise to the curation task. But sustainability is always an issue for such disciplinary services, and many if not most disciplines have never even got to the point where sustainability has to be confronted.

Conclusions

Australia appears at the point of devising a national framework for data services; this is highly desirable and should be strongly supported (Buchhorn and McNamara, 2006). In the US, the National Science Foundation Office for Cyberinfrastructure has just published its strategy (NSF, 2007). In the UK, there is little consistency across the Research Councils, with Arts and Humanities (AHRC), Social Science (ESRC) and Natural Environment (NERC) all having long-standing data deposit policies, and most of the others moving that way.

References

BATTERHAM, R., STANLEY, F., FROST, R., TSOI, A. C., FINNEY, K., CATHRO, W., KOTAGIRI, R. & ARTHUR, E. (2006) FROM DATA TO WISDOM: Pathways to Successful Data Management for Australian Science Canberra, Prime Minister's Science, Engineering and Innovation Council (PMSEIC) Working Group on Data for Science. http://www.dest.gov.au/NR/rdonlyres/D15793B2-FEB9-41EE-B7E8-C6DB2E84E8C9/15103/From_Data_to_Wisdom_Pathways_data_man_forAust_scie.pdf

BETHKE, S. (2000) Determination of the QCD coupling α_s *J. Phys. G: Nucl. Part. Phys.,* 26. http://www.iop.org/EJ/abstract/0954-3899/26/7/201

BOSE, R. (2006) Annotating Scientific Databases. Glasgow, University of Edinburgh, DCC. http://homepages.inf.ed.ac.uk/rbose/presentations/20060216_rbose_Glasgow_talk.pdf

BOSE, R. & FREW, J. (2005) Lineage retrieval for Scientific Data Processing: a Survey. *ACM Computing Surveys,* 37. http://portal.acm.org/ft_gateway.cfm?id=1057978&type=pdf&coll=&dl=ACM&CFID=15151515&CFTOKEN=6184618

BUCHHORN, M. & MCNAMARA, P. (2006) Sustainability Issues for Australian Research Data: the Report of the Australian e-Research Sustainability Survey Project. Canberra, Australian National University. http://hdl.handle.net/1885/44304

BUNEMAN, P. (2006) How to cite curated databases and how to make them citable. *Proceedings of the Conference on Scientific and Statistical Database Management.* http://homepages.inf.ed.ac.uk/opb/homepagefiles/harmarnew.pdf

BUNEMAN, P., CHAPMAN, A. & CHENEY, J. (2006) Provenance Management in curated databases. *2006 ACM SIGMOD international conference on Management of data.* Chicago, IL, ACM. http://portal.acm.org/citation.cfm?id=1142473.1142534

CARLSON, S. (2006) Lost in a Sea of Science Data: Librarians are called in to archive huge amounts of information, but cultural and financial barriers stand in the way. *The Chronicle of Higher Education.* http://chronicle.com/free/v52/i42/42a03501.htm

CHANG, G., ROTH, C. B., REYES, C. L., PORNILLOS, O., CHEN, Y.-J. & CHEN, A. P. (2006) Retraction of Pornillos et al., Science 310 (5756) 1950-1953. Retraction of Reyes and Chang, Science 308 (5724) 1028-1031. Retraction of Chang and Roth, Science 293 (5536) 1793-1800. *Science Magazine,* 314. http://www.sciencemag.org/cgi/content/full/314/5807/1875b

COLES, S., FREY, J. & MILSTED, A. (2006) Curation of Chemistry from Laboratory to Publication *UK e-Science All Hands Meeting 2006.* Nottingham. http://eprints.soton.ac.uk/41794/

DE WAARD, A. & VAN OOSTENDORP, H. (2006) Development of a Semantic Structure for Scientific Articles http://www.cs.uu.nl/people/anita/papers/deWvanOWIG2710.pdf

GALPERIN, M. Y. (2007) The Molecular Biology Database Collection: 2007 update. *Nucleic Acids Research,* 35**,** D3-D4. http://nar.oxfordjournals.org/cgi/content/abstract/35/suppl_1/D3

ICMJE (2006) Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication. IN EDITORS, I. C. O. M. J. (Ed.). http://www.icmje.org/

INTERPARES (2006) Appraisal of Electronic Records: a Review of the Literature in English. InterPARES Appraisal Task Force. http://www.interpares.org/documents/interpares_ERAppraisalLiteratureReview.pdf

IUCR (2002) Crystallographic Information Framework Version 1.1 Working Specification. International Union of Crystallography. http://www.iucr.org/iucr-top/cif/spec/version1.1/index.html

KERSEY, P. & APWEILER, R. (2006) Linking publication, gene and protein data. *Nature Cell Biology,* 8. http://www.nature.com/ncb/journal/v8/n11/full/ncb1495.html

MURRAY-RUST, P., RZEPA, H. & ZHANG, Y. (2004) Googling for INChIs; A remarkable method of chemical searching. *W3C Workshop on Semantic Web for Life Sciences.* Cambridge, MA, W3C. http://lists.w3.org/Archives/Public/public-swls-ws/2004Oct/att-0019/

NSB (2005) Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. IN FOUNDATION, N. S. (Ed.) *National Science Board Report.* http://www.nsf.gov/pubs/2005/nsb0540/

NSF (2007) Cyberinfrastructure Vision for 21st Century Discovery. Arlington, Virginia, National Science Foundation. http://www.nsf.gov/pubs/2007/nsf0728/index.jsp

PANG, K., YAU, K. & CHOU, H.-H. (1995) The Earth's Palaeorotation, Postglacial Rebound and Lower Mantle Viscosity from Analysis of Ancient Chinese Eclipse Records. *Pure and Applied Geophysics,* 145**,** 459-485.

PATRIAS, K. (2001) National Library of Medicine Recommended Formats for Bibliographic Citation. Supplement: Internet Formats. IN HEALTH, N. I. (Ed.). http://www.nlm.nih.gov/pubs/formats/internet.pdf

ROSS, S. (2000) Changing Trains at Wigan: Digital Preservation and the Future of Scholarship. IN OFFICE, N. P. (Ed.) *NPO Preservation Guidance Occasional Papers.* London. http://eprints.erpanet.org/45/

RUSBRIDGE, C. (in press) Tomorrow, and tomorrow, and tomorrow: poor players on the digital curation stage. IN EARNSHAW, R. (Ed.) *(Festschrift for Reg Carr).* Springer.

SHOTTON, D. (2006) The nature of biomedical research data. *Getting the most out of data, Making the most of research* London, Research Information Network. http://www.rin.ac.uk/files/David%20Shotton.pdf