Comments from the Digital Curation Centre on *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*, a draft report of the National Science Board. [1]

The development of the Internet has caused revolution in the way in which all forms of scholarship are conducted. The Scientific Method itself is changing as a result of the dependence of all sciences on their accumulations of digital data. The National Science Board report on long-lived digital data collections is a strong confirmation for the need to have policies, technology and methodologies for caring for – curating – these digital resources. The Digital Curation Centre (DCC) is an organisation supported by UK research and university councils, which is dedicated to these issues. The following comments[2] on the report are a summary of discussions among members of the DCC.

The DCC welcomes the draft report. In particular it endorses the view that curation is much more than preservation;:it involves all aspects of data acquisition and selection; it is about adding value to and organising data; and it implies *communication* both with current and future users of our digital resources. It also welcomes the use of "digital data collections" in preference to "databases" to emphasise the infrastructure and human effort involved in the curation process. It is in the management process where some of the challenging cultural and disciplinary differences arise; and these are as important as the technical challenges.

The DCC offers the following comments, which should not be interpreted as negative criticism but as additional issues that could be addressed in such a report. They are grouped, loosely, under the recommendations of the NBS report.

Develop a clear technical and financial strategy. Recommendations 1 & 2.

New technologies There are many aspects of digital curation for which technical advances are needed. The scientific world, for example, is making increasing use of databases for its digital substrate, yet the current state of database technology is not properly matched to the needs of digital curation. Important issues include the following.

- The need to reconcile collections that overlap in content but may be presented in differing data models or formats. This is often called data integration or semantic interoperability.
- The requirements of data exchange and the efficient export of data from disparate data collections.
- Annotation of data. It should be noted that some of the most valuable collections consist largely of annotation.
- Recording and tracking the provenance or origins of data. This is an important aspect of data quality, which is recognised by the NSB report as requiring further research.
- Archiving "dynamic data". Many data collections and most databases are not static. How do we preserve a data collection that constantly evolves during its lifetime?
- Security and protection. Almost all data collections have some fine-grain notion of access/update privileges; and these change over time.
- Transport and location of very large data sets. It is costly to move very large data sets. How does one distribute data and software so that computation remains feasible?

[1] `http://www.nsf.gov/nsb/meetings/2005/LLDDC_draftreport.pdf`

[2] This commentary is a summary of discussions of members of the UK Digital Curation Centre and their colleagues: Rajendra Bose, Peter Buneman, Peter Burnhill, James Cheney, Floris Geerts, Anastasios Kementsietsidis, Liz Lyon, Mags McGinley, Robin Rice, Chris Rusbridge, Charlotte Waelde and Stratis Viglas.

Of these only the first and last are receiving substantial attention by the research communities – partly as a result of Grid/Cyberinfrastructure funding.

We suggest that an additional recommendation would designate specific research and development areas for investment. The National Science Foundation is one of the organisations that could sponsor research on relevant topics.

Funding and the economics of long-lived data. The NSB report cites a number of well-established data collections. It should be remembered that several of these started off as small-scale efforts, and there are some highly-valued collections that remain relatively small-scale. The following points are relevant:

- There are several economic models for the maintenance of data collections. Which model is appropriate may change during the lifetime of the collection. This should be taken into account in planning both the development of data collections and the development of facilities to support them.
- There is a need for regular re-appraisal and peer review of data collections.
- Although the NSB report is addressed primarily to the National Science Foundation and US agencies, policy and technology recommendations will be much more effective if they have international acceptance.

We suggest that it is worth developing, perhaps as a part of recommendation 2, a comprehensive framework of requirements for management of long-lived data collections. Issues such as the economic model, requirements for preservation, and maintaining archival/evidential value would be covered in this framework. Not all collections would require all elements of this, but it would be a useful reference or base-line. Components of such a framework could include the OAIS reference model and a formalism of Representation Information".

Creating policy for key issues. Recommendations 3 – 6

Community-proxy functions. Recommendation 3 raises some questions around the costs of supporting "community-proxy" functions. and this is expanded in Section IV (6). The cost implications for institutions implementing digital commons infrastructure should not be under-estimated. Recent experience in the UK where many organisations have implemented "institutional repositories", suggests that although the set-up costs may not be great, the staff time (and therefore costs) in championing the approach and in ensuring that the repository is populated and maintained, are non-trivial. This cost has often been met by the university library or information service. Whilst most effort to date has not focused on the deposit of data collections, if one assumes that most of these collections will be in the "research data" category, and are distributed in departmental, group or personal archives, there is a significant area to be addressed. The debate as to whether these costs should be allocated to direct or indirect costs as part of the funding programme needs to be thoroughly examined by all funding agencies. In addition, there are likely to be legal and IPR issues, especially with data that has been generated as a result of consortium initiatives.

Research proposals that generate digital data. Recommendation 4 should be expanded to ensure NSF considers the incentives for deposit (and/or penalties for non or incomplete deposit, e.g. missing metadata) of data and the required metadata/documentation in collections, and that the use of existing data (compared with gathering new) should be an assessment factor in project selection. This recommendation should also recognise that neither the form nor content of the data generated by a speculative research project may be apparent at its inception, even though it is almost certain to generate useful data.

Training and cultural development. The NSB report introduces the broad term "data scientists" to describe scientists who are both familiar with some scholarly or scientific domain and who have expertise in the management of data collections. The relevant training and cultural development is discussed in recommendations 5 and 6. While provision for training of such experts is

laudable, elements of this training needs to be addressed much earlier in a researchers education. The creation, management and use of data collections should be introduced into the school (high school) curriculum, and continued as part of scientific training at the college level. The impact and costs of this change need to be addressed now.

Librarians have and archivists developed important skills in the collection, management and appraisal of data. However, their usefulness is sometimes questioned in the changing information landscape. This is an opportunity for these professionals to take on new roles, first through training in new information technologies and second by creating new forms of partnership with scientists and researchers. The development of these hybrid skills is an urgent requirement for educational and research organisations.

Intellectual property. The NSB report refers to the need to comply with intellectual property (IP) legislation. However, it does not address the question of whether this legislation is appropriate for the kinds of sharing required for the advancement of scientific knowledge. While IP can be a catalyst in the creation of scientific data, it can also be a powerful inhibitor – especially in the formation of new data collections derived from mixed sources. The report acknowledges that the legal issues connected with IP are complex and vary between countries. There are no easy answers here but some form of multi-national consulting or advisory body would be extremely useful.

Summary

This commentary has attempted to augment the recommendations of the NBS report. Several of the suggestions in this commentary stem from the need to consider further some of the technical aspects of digital curation and from the international nature of many digital collections. The authors are happy to enter into further discussions.

Peter Buneman
DCC Research Director
University of Edinburgh
opb@inf.ed.ac.uk

Liz Lyon
DCC Associate Director
University of Bath
E.Lyon@ukoln.ac.uk

Chris Rusbridge
DCC Director
University of Edinburgh
crusbrid@staffmail.ed.ac.uk