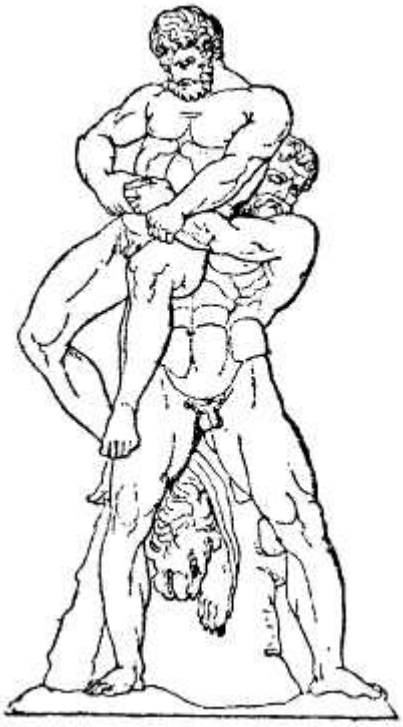
Strathprints Institutional Repository

The Antaeus Column*:
Data preservation, the new science and the practitioner librarian



Heracles defeating Antaeus.
Public domain image: from the old Swedish encyclopedia *Nordisk familjebok.*

* The title of the 'Antaeus' column derives from the name of the mythical giant, Antaeus or Antaios. The son of Gaia (whose name means 'land' or 'earth'), Antaeus was undefeatable in combat so long as he remained in contact with the earth. Once grounded by contact with the soil, he vanquished all opponents. However, in order to disempower Antaeus, Heracles simply lifted him from the earth, overcoming him totally. Thus, many times through the centuries, Antaeus has been used as a symbolic figure showing how any human aspiration must remain grounded in order to succeed. LIS research must therefore retain its contact with the 'ground' of everyday practice in order to fulfil its potential as a sophisticated research discipline – it must remain empowered by its relevance to practitioners.

Data preservation, the new science and the practitioner librarian

Abstract

Purpose of this paper	This paper outlines the information management principles of the so-called 'new science', and attempts to put these in the context of traditional library and information science principles.
Design/methodology/approach	A brief review of some work in the area, in particular focussing on the work show-cased by the annual digital preservation conference series hosted by the Digital Curation Centre in Scotland (http://www.dcc.ac.uk/).
Findings	There is a danger that scientists (as opposed to LIS professionals) will apply the information management techniques of the new science to their own activities inappropriately, especially to research that is best curated as 'old' not new science. This is something on which information professionals are well placed to give advice and make judgements.
Research limitations/ Implications	More practice-oriented research is needed to enhance understanding of how traditional librarianship practices can be applied to the data intensive scientific research carried out by so-called 'virtual organisations'.
Practical implications	This paper makes some initial suggestions about the how the tools of library and information practice can be related to the 'new science'. In particular, it highlights their relevance to distinguishing between the information management needs of the 'old' and the 'new' sciences: these needs are quite distinct, though easily confused.
What is original/value of the paper?	This paper relates terms from pure science such as the virtual organisation, cyberinfrastructure and e-science to traditional LIS concepts, and tries to create an understanding of the relationship between the two disciplines for the library practitioner.

Paper type: General review

Keywords: Libraries; digital preservation; cyberinfrastructure; e-science.

Introduction

Digital technology, as we know, challenges the Library's age-old role in collecting and preserving the documentary record of our civilisation's cultural and scientific output. For example, modern science works in ways that undermine the traditionally all-important role of journal-based scientific publication. Scientific experiments are so sophisticated that they are conducted in ways that defy expression in the formats (scientific papers) that libraries rely on as the building blocks of their collections.

Should librarians intervene in these digital preservation dilemmas and adapt their traditional expertise to solving this problem? Or should we stick to our traditional areas of print expertise and accept that the creation and preservation of digital knowledge is someone else's problem?

The problem of preserving our society's digital (rather than print) memory has become the subject of sophisticated academic study as well as the focus of innovative practical investigation. But what does all this mean for the practitioner librarian?

Libraries and the new science:

Data curation supplants the bibliographic record

Firstly, the LIS profession needs to think carefully about the trend whereby the significance of the bibliographic record as the core output of scientific experimental research may be diminishing. This is obviously a big challenge for us, since historically we have always assumed that to collect scientific research is to collect the definitive bibliographic record of it, the scientific paper.

This importance of this trend emerged strikingly from the first keynote at last year's International Digital Curation Conference (given by Hans Hoffmann[1], of CERN[2]). Increasingly, scientists value their raw data more than the bibliographic expression of that data, and view the preservation of raw data as the prime curational challenge for the knowledge professions such as librarians and archivists. And if these knowledge professions can't help, then scientists will seek a solution elsewhere, from a new generation of expert digital curators and data preservation technicians!

In the past, the object of science was to collect data in order to generate general rules or summary findings. These findings could be distilled into scientific journal articles which carried condensed narratives of relevant experiments. These accounts acted as the 'tip of the iceberg' for spreading scientific knowledge. The raw empirical data was often lost – but that was ok, because the tip mattered more than the iceberg. Subsequent experiments conducted along the lines of the summary findings in papers would confirm the empirically verifiable repeatability of the previous findings. And in turn the complete, detailed raw data of subsequent experiments was not reproduced in such follow-up work.

However, the new science is different. New science is distinguished by large scale, costly experimentation where it is difficult or impossible to exhaust the data produced in a brief series of papers published simultaneous with the progress and conclusion of the experiment. The cost of such large-scale experimentation is also so great that, having created the data, the costs of repeating the experiment to

generate the same data would be prohibitive. So the data itself is the most important outcome of the experiment.

The structures needed to support and maintain such science is called 'e-science' in Europe and 'cyberinfrastructure' in the USA[3] (and to show that this is something that calls for a library reaction to its implications, see the evidence of the US library profession's attempts to confront it at http://www.arl.org/forum04/). This science is typically conducted by large, geographically dispersed, but virtually connected organisations – 'Virtual Organisations'. Although these terms seem to be conference buzzwords, they are more than that and do have a clear meaning which is genuinely useful. The impact on the library, archival and curatorial professions is traced in a variety of publications[4].

First catch your quark

In Europe the long-standing high energy physics collaboration based at CERN[2] in Switzerland is the biggest and best known example of experimentation that is impossibly expensive to repeat, and which generates mountains of data that will take decades to fully interpret.

Of course, we, as naïve onlookers, would probably liken CERN's discovery of the existence of fundamental particles of matter to the discovery of an unknown species of fauna. The recent discovery of a colossal squid[5] in the Antarctic is our vision of scientific discovery – the pictures and documentation of the discovery will be a lasting record, long after the frozen animal has rotted down and melted back into the sea.

But this is old science. The discovery of quarks, hadrons and charm in a particle accelerator is not like that. Whereas the Antarctic fisherman posed excitedly with their big squid, Hans Hoffman at no point emerged from a particle accelerator holding a quark up to be photographed. Rather than being 'determinant entities' that can be stuffed and mounted in a museum, fundamental particles are 'information events' – slide 6 of Hoffman's CERN presentation[1] is his nearest attempt to photographing a fundamental particle. What we see is a grid with numbers on it, some of which are circled. There is an arrow pointing to the circle saying 'Interesting physics!' What is inside the circle is a data pattern, not a 'thing' as such – the particle that generated the data pattern died more or less as soon as it was born. In fundamental physics, when you pull your nets in, the creature of the depths has disappeared!

So quarks really don't look like much at all: this is the new science. This process of discovery sounds like good news for information professionals – if the objects of such enquiries are not realia but information, then surely we are ideally placed to store such information? Not necessarily: librarians collect and preserve documents which are themselves interpretations of data. Pure data curation is new to us.

What is the role of the library profession?

We can at least say that this is something on which we have to formulate a view. The old science enshrined interpretations of data as bibliographic outputs as journal articles which have been long familiar to us. Do we now want to be responsible for the storing of data prior to such interpretation, in such a way that its integrity is not compromised? Does the LIS profession possess sufficient knowledge of science to enable it to make such decisions? Would librarians without PhD qualifications in High Energy Physics recognise any such compromised data if they saw it?

For example, if data has to be stored, it must be moved from platform to platform, as platforms decay and new ones replace them. Are the technical and subject skills need for such migrations best left as the arcane responsibilities of the scientists who generated the data in the first place, or are there, at the heart of it, new generic skills which can be integrated into the library tradition of knowledge preservation?

These are obviously big questions which institutions like the UK's Digital Curation Centre will research and explore for us, in the hope of pointing a way forward. But practitioner librarians should not lose their nerve when confronted by these questions. If the preservation of knowledge were to become the unique preserve of those who alone understand its fullest complexities, then the possibility of shared knowledge itself disappears. The scientific community will itself break up into silos, a series of knowledge islands where solipsistic ghetto-dwellers treasure the sacred data of their grim little tribe.

In particular there is a danger that scientists themselves will apply the information management techniques of the new science to their own activities inappropriately. This is something that information professionals are much better placed to judge and give advice on, than (with respect) information-naïve scientists. Scientists who are very much 'within' their own data, have little overview and sense of perspective as a result of this inevitable tunnel vision.

A general scan through some of the recent work on digital preservation and data curation does unfortunately show this trend – though the reader can skim the papers at the 2006 Glasgow international conference on data curation and its 2005 predecessor in order to make up their own mind about this[6]. Without being too specific, a jobbing practitioner librarian would probably say that some quite modest scientific research projects are – arguably - using the information techniques of the new science inadvisedly.

A particular problem is the creation of exotic project-specific taxonomies because the researchers feel their data is so special that it needs this sort of treatment. CERN-like projects carried across vast Virtual Organisations (VOs) may need to create new internal taxonomies because of the uniqueness of the materials they are exploring, because then their data can be shared meaningfully across the VO, and curated and preserved with integrity into the future across evolving data storage platforms.

However, there is a lot to be said for just using off the shelf taxonomies that LIS professionals have used for many years. After all, they are known to other knowledge professionals, most intelligent scientists and information users (these are the merits of standard library classification schemes). In this way researchers outside the VO might have a better idea of what the VO's research data means and they can help integrate the data into the general pattern of shared research.

So traditional librarianship has a lot to offer the new science.

At the very least, the traditions of the LIS profession are a metaphor for the need to share fundamental scholarly knowledge. The fact that the LIS profession has for millennia given our society a shared cultural and scientific knowledge base, one that can be ordered and preserved in terms of generic structures open to and understood by all, such as general classification schemes, user friendly catalogues, subject bibliographies and indexes, is a fact to treasure, a source of deep professional pride. Our wish to maintain this role is not professional self-interest: it is simply an expression of our awareness that, for the information society to work, information

must be held in common. Librarians facilitate this by designing the information structures that underpin the process of knowledge exchange - and they can do this as much for the new science as they have always done for the old.

Nicholas Joint,
Centre for Digital Library Research/Andersonian Library,
University of Strathclyde.

References

1. Hoffman, Hans. (2006) Opening keynote of the 2nd International Digital Curation Conference 2006. Digital Data Curation in practice.
< http://www.dcc.ac.uk/events/dcc-2006/programme/presentations/h-hoffmann.ppt > Accessed 29/3/07.

2. CERN: Conseil Européen pour la Recherche Nucléaire, or European Council for Nuclear Research. < http://public.web.cern.ch/Public/Welcome.html > Accessed 29/3/07.

3. "The term e-Science (or eScience) is used to describe computationally intensive science that is carried out in highly distributed network environments, or science that uses immense data sets that require grid computing. The term was created by John Taylor, the Director General of the United Kingdom's Office of Science and Technology in 1999 and was used to describe a large funding initive starting in November 2000. Examples of the kind of science include social simulations, particle physics, earth sciences and bio-informatics. Particle physics has a particularly well developed e-Science infrastructure due to their need for adequate computing facilities for the analysis of results and storage of data originating from the CERN Large Hadron Collider, which is due to start taking data in 2007." (Wikipedia)

4. Goldenberg-Hart, Diane (2004) Libraries and Changing Research Practices: a report of the ARL/CNI Forum on E-Research and Cyberinfrastructure, *Libraries and Changing Research Practices.*
< http://www.arl.org/newsltr/237/cyberinfra.html > Accessed 29/3/07.

5. (Anon) Colossal Squid Caught off Antarctica. National Geographic, February 22, 2007.
< http://news.nationalgeographic.com/news/2007/02/070222-squid-pictures.html?source=G1902 > Accessed 29/3/07.

6. 1st and 2nd International Digital Curation Conferences, Bath/Glasgow, UK, 2005/2006.
< http://www.dcc.ac.uk/events/dcc-2005/ and http://www.dcc.ac.uk/events/dcc-2005/ > Accessed 29/3/07.

Received 29/3/07, reviewed 30/3/07, revised 31/3/07, accepted 1/4/07.