Digital curation at Gloucestershire Archives

Viv Cothey

Gloucestershire Archives, Clarence Row, Alvin Street, Gloucester, GL1 3DW.

viv.cothey@{gloucestershire.gov.uk, wlv.ac.uk}

Abstract

There is a need, particularly within the local government sector, to implement prototype or demonstrator software tools in order to develop digital curation practice. Conceptional frameworks relating to digital curation, for example the reference model for an Open Archival Information System, are now well known. Less well known are developments in workflow tools. In this presentation I shall consider digital curation practice in the higher education sector and the scientific research community. The focus and direction of these developments will be compared with developments in digital curation practice at Gloucestershire Archives. The presentation will also touch upon some of the challenges experienced when advocating and implementing digital curation action.

Introduction

There really is quite a lot going on as regards digital curation. In the UK the JISC[1] *Information environment programme (2009-2011)* continues the excellent work carried out under the *Repositories and preservation programme (2006-2009)* and before that the *Digital repositories programme (2005-2007)*. In the US there is the National Digital Information Infrastructure and Preservation Program,[2] while in the EU we have challenge four of Framework 7: "*Digital libraries and content*"[3] and Australasia has ICADS/PADI.[4]

The DPC[5] and others including several international conference series as well as a handful of specialist journals monitor the digital curation world and keep us informed.

If I fail to mention your favourite organisation/conference/journal I apologise in advance; there will certainly be very many groups doing good work that I ignore here.

So what is the problem?

Looking out on the world from the perspective of a local authority record office there seem to be two problems. Firstly, digital curation appears to be the exclusive province of "the big players", or certainly players that are bigger then county record offices! In particular it

[1] Joint Information Systems Committee, `<http://www.jisc.ac.uk/>`.

[2] See `<http://www.digitalpreservation.gov/>`, I include here RLG/OCLC and Fedora/DSpace also.

[3] See `<http://cordis.europa.eu/fp7/ict/programme/challenge4_en.html>`.

[4] Preserving Access to Digital Information, `<http://www.nla.gov.au/padi/about.html>`. STORS in Tasmania is also noteworthy `<http://www.stors.tas.gov.au/about/>` as is Greenstone `<http://www.greenstone.org/>`.

[5] Digital Preservation Coalition, `<http://www.dpc.org>`.

seems to be commonplace that there is access to specialist IT resources. Second, there seems to be an absence of practical end user systems that might provide low cost learning opportunities for archive professionals. In consequence, I suggest, local authority archivists are largely “out of the loop” when it comes to digital curation. Local authority archivists may well be fully aware of the very many exhortations[6] to do digital curation and to get involved but are frustrated by not knowing where to start.

Like many of us, my colleagues are pressed for time and have little opportunity to learn about digital curation. In any event there seem to be so many prerequisites and even obstacles to learning about digital curation. Not least there is a new digital curation jargon which is used to express crucial technological concepts that previously one might have safely ignored. Is it easy or difficult to explain to the naive why digital curation is more than either,

a) storing recordable CD/DVDs in a chilled strongroom, or
b) taking daily backups of the council's computer network servers?

Providing explanations soon draws one into conversations about OAIS,[7] proprietary file formats, digital libraries, web archiving, digital rights and so... . And then one is asked, “but what can we actually do?”

It was against this background that at Gloucestershire Archives we developed *GAip*.

GAip

Gloucestershire Archives ingest packager (GAip) is a prototype desktop workflow tool. The concept of an archival ingest package (AIP) was introduced in the OAIS which has now been elevated to the status of an international standard (ISO 14721:2003). BagIt[8] (Boyko *et al*., 2008) is an example of a digital package. GAip provides the means to automatically compile a BagIt like archival ingest package. This is carried out in a desktop computing environment so that an archivist can curate, e.g. package, a digital object[9] or collection of digital objects at the click of a button.

Whilst GAip provides useful curation functions, its principal role is to support hands-on experiential learning. That is, GAip users learn about digital curation by doing.

A secondary benefit of GAip is to provide credibility when advocating the introduction of dgital curation practice. Even though, as yet, best practice in digital curation falls somewhat short of providing complete solutions, this does not justify avoiding action and is definitely not an excuse for despair. Good curatorial action based on sound principles is already a practical reality.

Digital libraries

Much of the activity mentioned earlier was motivated by the recent quite rapid transition in academic libraries following the Follet Report (Higher Education Funding Council for England, 1993). The move towards electronic information delivery, particularly in respect

[6] Such as the DPC's seminal “*Mind the gap*” report (Waller and Sharpe, 2006) and subsequent literature.
[7] Open Archival Information System
[8] See `<http://www.digitalpreservation.gov/library/challenge/data-transfer.html>`.
[9] The generic term object is used here to include individual files, directories, zip archives etc.

of journals and then learning materials more generally, has had the desired effect of reducing the strain on physical shelfspace but has prompted concerns over the long term accessibility/availability of material. These concerns combine with concerns over the journals pricing model (which remain unresolved) to stimulate the growth in interest in so-called open access and so-called "self-archiving"[10] Institutionally based digital libraries have been implemented that support *self-accessioning*[11] by academic authors across a wide range of digital material. Indeed it is suggested by some[12] that author self-accessioning (and open access) should become the norm and be obligatory when research is publicly funded. Similar complementary developments in e-Science have highlighted the digital curation needs of scientific datasets.[13]

Institutional digital repositories are now an established part of the academic and scientific research sector. Examples include both broad scientific communities and particular specialities[14] as well as individual organisations. A recent milestone is the comprehensive provision of institutional digital repositories across all higher eduction in Wales[15] (Knowles and Lewis, 2009).

The maturity of the digital library/repository market is also reflected by a consolidation of early major players. For example Fedora and DSpace have merged to form DuraSpace.[16] It is probable that other well known names will either merge or disappear.

However there is also a growing realisation that digital libraries/repositories *per se*, despite appropriating adjectives such as *archive,* actually fall somewhat short of meeting the aspirations of archivists concerned with curating digital objects. The current "KeepIt" project looks to make up some of the shortfall.[17]

Although several digital libraries/repositories provide server specific client or user interfaces to support author activity, this is neither universal nor uniform. JISC identified the lack of a uniform interface to deposit digital objects as one of the obstacles to authors fully engaging with digital libraries/repositories. Accordingly they sponsored the development of the simple web offering remote deposit (or sword[18]) server-client function or protocol.[19] Many digital libraries/repositories have adopted the sword protocol so that an author can use a single client interface to deposit their digital objects in any[20] digital library/repository.

[10] But clearly not archives as we know it!

[11] The term "self-archiving" is promoted in favour of "self-publishing" which is deprecated because of its associations with vanity publishing. However I prefer "self-accessing" since this more accurately describes the library function.

[12] For example, see <http://openaccess.eprints.org/index.php?/archives/136-guid.html>.

[13] For example, see <http://wiki.ecrystals.chem.soton.ac.uk/images/8/82/ECrystals-WP4-PP-090625.pdf>. and <http://www.jisc.ac.uk/publications/documents/keepingresearchdatasafe.aspx>.

[14] For example <http://arxiv.org/> or Lyon et al, (2008).

[15] See also <http://www.jisc.ac.uk/whatwedo/programmes/reppres/sue/welshrepositorynetwork.aspx>.

[16] See <http://duraspace.org/index.php>.

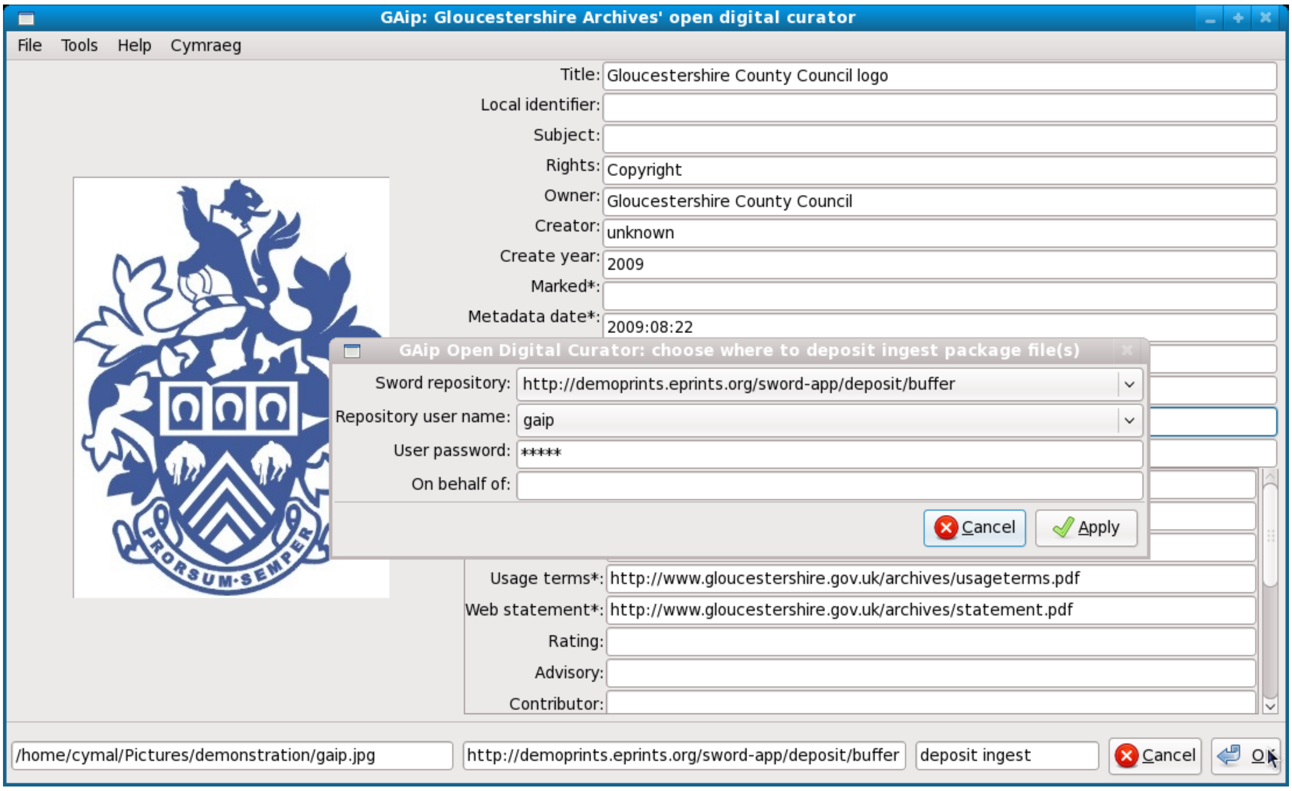
[17] See <http://www.jisc.ac.uk/whatwedo/programmes/inf11/keepit.aspx> and <http://preservation.eprints.org/keepit/>.

[18] Sword, <http://swordapp.org/>.

[19] See <http://www.jisc.ac.uk/whatwedo/programmes/reppres/tools/sword>.

[20] Assuming of course that the repository is sword compliant, see <http://swordapp.org/sword/implementations>.

CyMAL[21] are funding work by Gloucestershire Archives to include sword as part of GAip. In consequence an archivist using GAip will be able to deposit an ingested curation package in a digital repository of choice. Figure one illustrates a "screen-shot" showing the archivist's user interface.


GAip: Gloucestershire Archives' open digital curator
File
Tools
Help
Cymraeg
GAip Open Digital Curator: choose where to deposit ingest package file(s)
Cancel
Apply
Cancel
OK

Title: Gloucestershire County Council logo
Local identifier:
Subject:
Rights: Copyright
Owner: Gloucestershire County Council
Creator: unknown
Create year: 2009
Marked*:
Metadata date*: 2009:08:22

Sword repository: http://demoprints.eprints.org/sword-app/deposit/buffer
Repository user name: gaip
User password: *****
On behalf of:

Usage terms*: http://www.gloucestershire.gov.uk/archives/usageterms.pdf
Web statement*: http://www.gloucestershire.gov.uk/archives/statement.pdf
Rating:
Advisory:
Contributor:

/home/cymal/Pictures/demonstration/gaip.jpg
http://demoprints.eprints.org/sword-app/deposit/buffer
deposit ingest

Figure 1: Using GAip to deposit an ingest package

When the "OK" button is clicked GAip will create an AIP containing the image shown together with the metadata specified and will deposit it in the digital repository at `<http://demoprints.eprints.org/>`.

Integrated curation workflow

The notion of a fully integrated "scholar's workbench" that provides a close coupling of an author's creative desktop computing environment with a digital library is discussed by Barnes, (2007). The current Hydra project[22] is an implementation of these ideas (Green & Awre, 2009). The vision is that authors' born-digital material will exist in a digital repository and be managed from inception to publication and from there on to "preservation". Not surprisingly there is a strong similarity between this vision and the DCC[23] curation lifecycle model (Digital Curation Centre, 2009).

Yet in the same way that archivists with an emphasis on collection, context and provenance differ from the more item by item bibliographic approach preferred by

[21] See `<http://wales.gov.uk/topics/cultureandsport/museumsarchiveslibraries/cymal/?lang=en>`.

[22] See `<https://fedora-commons.org/confluence/display/hydra/The+Hydra+Project>`.

[23] Digital Curation Centre, `<http://www.dcc.ac.uk>`.

librarians, so the self-accessioning predicated by the scholars workbench may not meet the curatorial needs of digital archivists. An obvious difference is in the approach to cataloguing and resource discovery. Digital libraries lean upon, for example Library of Congress or “special library” classification schemes. It is believed that a UKAT[24] based classification scheme would be found more appropriate by local government archivists.

However, despite some reservations, it is clear that the academic sector's development of digital workflow tools to complement their digital library infrastructure is worth watching. It is also worth watching the ways in which the academic sector makes use of open source software[25] For many local authorities, including ours, supporting users' running open source is a novelty. This is much less so in higher education. Many of the digital workflow tools being developed are open source. Indeed, open source and especially open formats have an important role to play in digital curation.

Experience using GAip has informed its development. At the outset GAip's functionality was focussed on the gaip ingest package. GAip operations were limited to creating an ingest package, extracting a file from an ingest package and some editing of the package metadata. GAip is now becoming more of a general digital curation workbench in that it now operates on a range of digital objects, not just packages. In particular GAip now supports a form of “digital production”. Dissemination versions of digital images or documents can be automatically created that include relevant metadata, especially related to copyright. The sword enhancement mentioned earlier allows for depositing digital objects in a repository.

Trusted digital repositories

We have already seen that there are growing aspirations within the academic sector to provide “digital archive” quality preservation for digital objects. DPE's[26] planning checklist PLATTER[27] (DigitalPreservationEurope, 2008) sets out an operational framework for creating and running a trusted digital archive. In particular there is a focus on trust. It is not enough to to safeguard digital objects long term, one must demonstrate that this is being done.

Does the existing infrastructure of digital libraries provide trusted digital repositories in the PLATTER sense? If it does then what kind of “digital archivist's workbench” might be needed to front-end the digital library in order to facilitate curatorial activity?

The Society of Archivists, through their research fund, are sponsoring Gloucestershire Archives to investigate a particular instance of an academic digital repository from the point of view of local government archivists interested in providing a trusted digital repository service. This project will run from September 2009 to March 2010.

Conclusion

Some of the technical and infrastructural challenges posed by the needs of digital curation are big and some solutions will take a long time to develop. But this does not have to be a

[24] UK archival thesaurus, <http://www.ukat.org.uk/>.

[25] See for example, <http://www.oss-watch.ac.uk/resources/opensourcepolicy.xml>.

[26] Digital Preservation Europe, <http://www.digitalpreservationeurope.eu>.

[27] Planning Tool for Trusted Electronic Repositories

justification for procrastination.

On the contrary, there is a current window of opportunity for local government archivists to explore small-scale, low cost, prototypical digital curation in order to develop skills and awareness. This is what Gloucester Archives is doing.

In so doing we also hope to make a contribution not only to encourage appropriate action within our funding council (and discourage inappropriate action) but also to support progress in digital curation across the sector.

References

Barnes I. (2007). The digital scholar's workbench. In *Proceedings of the 11th International Conference on Electronic Publishing* eds. Chan and Martens, pp. 285-296. Vienna.

Boyko A., Kunze J., Littman J., Madden L., and Vargas B. (2008). *The BagIt file package format*. Available from: <http://www.digitalpreservation.gov/library/resources/tools/docs/bagitspec.pdf>. [Accessed: 23 August 2009].

DigitalPreservationEurope (2008). *DPE Repository Planning Checklist and Guidance*. [Project 034762]. Available from: <http://www.digitalpreservationeurope.eu/publications/reports/Repository_Planning_Checklist_and_Guidance.pdf>. [Accessed: 23 August 2009].

Digital Curation Centre. (2009). *Curation lifecycle model*. Available from: <http://www.dcc.ac.uk/lifecycle-model>. [Accessed: 23 August 2009].

Green R. and Awre C. (2009). *T*owards a repository-enabled scholar's workbench: RepoMMan, REMAP and Hydra. *D-Lib magazine,* 15(5/6).

Higher Education Funding Council for England (1993). *Joint Funding Council's Libraries Review Group: Report,* HMSO.

Knowles J. and Lewis S. (2009). *WRN final report*. Available from: <http://ie-repository.jisc.ac.uk/313/>. [Accessed: 23 August 2009]. JISC.

Lyon I., Coles S., Duke M. and Koch T. (2008). *Scaling up: towards a federation of crystallography data repositories*. Available from: <http://www.ukoln.ac.uk/projects/ebank-uk/dissemination/Ebank3report/Ebank3report.pdf>. [Accessed: 23 August 2009]. UKOLN.

Waller M. and Sharpe R. (2006). *Mind the gap: assessing digital preservation needs in the UK*. Digital Preservation Coalition.