The International Journal of Digital Curation
Issue 2, Volume 2 | 2007

Report from the DCC Workshop: Legal Environment of Digital Curation

Angus Whyte,
Digital Curation Centre, University of Edinburgh

November 2007

Summary

This is a report from the Legal Environment of Digital Curation workshop held at Glasgow University on November 23, 2007. The event provided an overview of legal considerations for non-legal professionals who work with data, focusing especially on intellectual property rights and licensing, data protection, freedom of information and privacy, and data as evidence. The workshop was organised in conjunction with the *SCRIPT-ed* journal of law and technology, and supported by JISC, the AHRC and Edinburgh University.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

SOME RIGHTS RESERVED

Introduction

Billed as an 'overview of legal considerations for non-legal professionals who work with data' this event coincided with a furore over the loss of confidential data by the UK Government's Revenue and Customs, the latter reportedly prompting the Information Commissioner to remark that "it just does not matter what laws, rules, procedures and regulations are in place, if there is no proper enforcement of those rules." (*The Guardian*, 2007)[1]. The workshop was a useful opportunity for those charged with keeping a watchful eye on both this and the less obvious risks and opportunities that the legal environment presents data professionals, including database licensing and other intellectual property rights (IPR) issues.

Andrew McHugh, the DCC's Advisory and Audit Services Manager, welcomed the participants. He pointed to the all-encompassing relevance of legal issues; as we face the numerous challenges of adding value to data; the digital curation field has seen more acknowledgment of issues of organizational sustainability, including IPR issues, and methodologies and metrics to assure the legitimacy of digital data repositories. Legal issues have been at the fore in his work with repositories, common concerns being 'what can we do' in licence terms and 'what should we do' to ensure compliance.

Mags McGinley introduced the opening session with "a legal perspective on digital curation". She began with a question: What areas of law are relevant to digital curation? She included Intellectual Property Rights, Information Governance, Evidence, Contract, Accessibility, Defamation, Human Rights, Legal Deposit, and Security. Her list was not exhaustive however, because the answer is as broad as the range of what counts as "data". We should consider any aspect of law that places data in an institutional context. That includes the management of digital objects, organizational and technical infrastructure and security.

From that broad view, the three main issues of the day proved to be of widespread and current concern. Curation depends for its sustainability on viable **IPR and licensing models**. Robust curation practices will help an organisation comply with **Data Protection and Freedom of Information** statutes, and bring management of digital resources in line with paper where good practice tends to be better established. **Evidential value and authenticity** are important from two angles. Firstly a mandate for curation will stem from the importance of data as legal or scientific evidence, e.g. as a record of non-repeatable observations. Secondly for accountability purposes, cost-effective measures are needed to prove what has been done, and that it has been done correctly.

Graham Pryor, DCC eScience Liaison Officer, followed with an overview of the two inter-locking hemispheres of eScience: research infrastructure and the research it supports. Two exemplars of these were, respectively, the StORe[2] project's work to link publications with the datasets they refer to and understand the ramifications of doing so; and the CARMEN consortium[3]. The latter brings together neuroscientists and

[1] DCC Blawg provides commentary on this and other current legal issues affecting curation http://dccblawg.blogspot.com/
[2] StORe. (Source-to-Output Repositories) http://jiscstore.jot.com/WikiHome
[3] http://www.carmen.org.uk/

informaticians from 19 partner organizations, illustrating the emergence of hybrid fields, interdisciplinary aims, and the aim of creating new knowledge using the new infrastructure. Another two-part way of seeing e-Science is as input and output. eScience takes an increasing proportion of the £3 billion annual research budget. It generates data that is large in scale and complexity. This transformation is engendering policy changes on data sharing and preservation from the funding bodies, ranging from gentle encouragement to 'carrots and sticks'.

While the technical development of eScience gathers pace, the legal environment is uncertain; linking data and publications, for example, raises legal and technical issues about protection from 'predators'. Pryor identified hotspots in the area of access to and ownership of IP. Where eScience lowers technical constraints on the ability to access remotely and analyse complex datasets, the corollary is a greater need for explicit policies on individuals' rights to do so, and to re-examine implicit assumptions about data ownership.

Intellectual Property Rights and Licensing

Two sessions on IPR highlighted relatively new aspects of a familiar topic; the copyright issues surrounding digital preservation, and the EU-wide legislation on IPR in databases.

Adrienne Muir of Loughborough University began by surveying IPR issues of special interest to digital archives and repositories. The terms repository or archive imply a commitment to the long-term, but that commitment needs to be backed up by a mandate from the relevant stakeholders. In particular, 'output' repositories need to persuade publishers that it is safe for them to allow programmes to archive and preserve. That in turn requires definition of rights and responsibilities. Contractual rights are needed to acquire, normalise if appropriate, store, preserve and provide access. Responsibilities need to be clear – who does what and who is responsible for what, and what happens if there are changes in ownership of content through transfers, mergers and acquisitions.

Services are at an early stage of definition, as Muir pointed out, so work on auditing and certifying digital archive services is key to establishing their legitimacy, as are newly emerged standards such as OAIS, METS, and preservation metadata. Auditing is needed to warrant the archive's administration, its organisational viability and financial sustainability; alongside its technological suitability, come system security and procedural accountability. Important groundwork can be found in the TRAC Criteria and Checklist (Center for Research Libraries [CRL], 2007b), the CRL Certification of Digital Archives project (CRL, 2007a), and the DCC pilot audit programme (DCC/DPE, 2007).

Archives may only carry out 'restricted acts' on copyright works if permitted under an exception to copyright law, which defines the otherwise exclusive rights of copyright owners to reproduce a work, issue copies to the public, rent or lend, perform, show or play the work in public, broadcast the work or make an adaptation of the work. The copyright owner also has the right to prevent third parties from carrying out these "restricted" acts without prior permission. UK copyright law specifies copyright duration as normally 70 years from the end of the year when the creator died or if the work is anonymous, pseudonymous or produced by a company, 70 years from end of

the year it was published (special rules for unpublished works). Copyright protection for some other works is set at 50 years from creation, including sound recordings, broadcasts and computer-generated works.

Muir went on to describe copyright issues relating to the various means of digital preservation; r**efreshment/media migration**, **migration** from one format to another, e**mulation** using software to enable a new technology platform to mimic an older one; or r**ecreation**, which creates a new digital object representing significant properties of the original, but does not incorporate any elements of the original digital object.

She highlighted possible "copying" requirements and issues that these approaches raise for the ingest, preservation and delivery of digital material. Firstly, ingest may require copying data from the original medium, and reformatting it. It may involve encapsulating the content, the original software and specifications etc., and extracting, reformatting and saving metadata for preservation. Secondly, preservation may require periodic copying of bit streams from one physical medium to another, periodic content format conversion, and recording and saving information about the original software environment. To render the content, an archive may then need to use emulation software, recreate a software environment, or create a dissemination format.

As well as the 'copying' implications, each stage may be affected by problems of copy-protected media, of losing the look and feel or even meaning of the data through conversion, and of the need to obtain and save documentation and metadata – manuals, specifications, etc. Furthermore, using emulation may be the equivalent to making an adaptation.

To compound the uncertainty, the exemptions for preservation that are made under the copyright legislation are oriented to paper publications. Libraries and archives can make copies for preservation purposes, if they are neither established nor conducted for profit, nor part of a body established or conducted for profit. If a library undertakes copying to replace a copy in another library, the materials must form part of the "permanent collection" of both the donor and the receiver libraries and must be for reference use only. The term "reference use" is reasonably clear in the print environment in that material may be consulted on the premises only and should not issued as a loan and taken off the premises. However this does not translate well to digital access! The litmus test for all such exceptions is the Berne Convention's "three-step" test, which states that they must be restricted to "special cases which do not conflict with a normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the rights holder."

In short, Muir concluded, archives cannot rely on copyright law exceptions as an effective basis for carrying out digital preservation. Archives must therefore obtain licences from rights holders. There is no blanket licensing scheme for digital copying for preservation purposes; although recently a model licence has been developed specifically for electronic journals for the academic sector , which provides for archiving in the event of subscriptions being terminated (NESLi, 2007). So until and unless current pressure to amend copyright exceptions succeeds, archives must do two things. Firstly, work out all the rights that will be needed to carry out digital preservation, and secondly, check if any licences used by rights holders provide the rights needed and/or secure these rights from the rights holders.

The second presentation on IPR was by legal consultant **Jordan Hatcher**, on **database rights and licensing**. Hatcher described what the database right is and how to get it, and why 'open data' licences may be the preferred approach for many. First he pointed out that the database right is additional to the copyright that is attributable to the data held. The database right, which like copyright covers copying, distribution and adaptation, is a "new" (*sui generis*) right that applies to the database itself - if there has been a substantial investment in obtaining, verifying or presenting its contents.

To qualify as a database, the data need not even be digital, provided it is the result of selection and arrangement and is the author's intellectual creation. Each entry "must be independent", and the arranged entries systematic and methodical. Hatcher pointed out that the right does not apply to software used to manipulate the data, but the line can be unclear.

The database right, Hatcher continued, is an automatic right and applies throughout the EU for 15 years from creation. It is automatically earned by the database creator and can also apply where the creator does not create the data itself - but does invest in verifying and presenting it (as in sports fixtures, for example). If the creator puts in further investment, the right can be extended indefinitely. The investment must be "substantial", although it is unclear what minimum applies to that.

The database right will need to be acquired from a database creator, unless the use made of the database is restricted to insubstantial extraction and re-use. Of course anyone with a collection may have database rights in it. However Hatcher pointed out that exercising the database right does not come without cost. That lies in the effort needed to negotiate licences for use and re-use.

Hatcher's key point was on the benefits of an 'open data' approach. The copyright and database rights legislation offers protection to database creators, in that others cannot re-use the data without permission. However, for both the database creator and users/re-users, this may be complex and time-consuming to obtain. Open data, on the other hand, applies similar principles to those found in open source software - fundamentally, that work should be available for use and re-use by the public without the need to seek further permission from the rights holder. A problem is that the Creative Commons licences do not cover database rights. Nor, Hatcher argued, is the GPL open source software licence appropriate for databases and data.

Hatcher is pursuing the approach through ongoing work on the 'Open Data Commons' set of licences, which are soon to be made available for comment (Hatcher, 2007). These take a "some rights reserved" approach, enabling database producers to give their intended users permission while still influencing their behaviour with the data/database. He identified some problem areas; moral rights, patents and trademark. For academic purposes, licences also need to take account of rights of first publication, attribution/citation, and the different norms across disciplines, these being difficult to assess and problematic to cover in a licence aiming for universality. Possible solutions are to use contracts in addition to licences, limit the reach of licence elements like attribution and share alike, or waive all rights altogether.

Overall, according to Hatcher the approach should benefit database consumers by clearing rights for use in advance, and benefit producers by lowering barriers to the use and re-use of their work.

Data Protection, Freedom of Information and Environmental Information

"When Worlds Collide" was the aptly named title of the first presentation in this session. **Renata Gertz** of the AHRC Research Centre in Edinburgh University's School of Law guided the audience through pitfalls of UK legislation enshrining two "diametrically opposed" principles, especially as applied to health data. The first principle is confidentiality, promoted by the Data Protection Act 1998, while the second principle is openness, as expressed in the Freedom of Information (FoI) acts. Recent court cases in Scotland and England have tied these principles in knots that Gertz attempted to unravel.

As is well known, the Data Protection Act protects 'personal data' from unlawful disclosure to third parties, while since 2005 FoI has given the public a general right of access to information held by or on behalf of public authorities. The troublesome interface between the two arises from the definition of 'personal data'. This refers to data relating to a living individual who can be identified from those data or, more problematically, from "those data and other information which is in the possession of, or is likely to come into the possession of, the data controller". This definition applies to both the Data Protection and FoI legislation, and the former includes health-related data in the class of "sensitive" data to which stringent security measures should be applied.

The problem Gertz described lies in the potential for disclosure of personal data arising from analysis of data collections that are released to comply with FoI. Following the 2006 case of the Common Services Agency vs Collie it is not currently clear what is legally permissible, as the Scottish Information Commissioner ruled that this health body was correct to regard data on childhood leukemia cases analysed by year and census ward as "personal". The ruling was on the grounds that the combination of rare diagnosis, a specified age group, and the small area meant that the numbers would be sufficiently small potentially to identify individuals. However this ruling was qualified; the data should be made available in "barnardised" form, i.e. ostensibly made less revealing by randomly adding or subtracting 1 from some cells in the table while leaving the total intact.

There are no absolute boundaries around the numbers of personal characteristics disclosed that are sufficient for data to count as personal; Gertz pointed out that this depends on how easily a person may be identified. However an additional layer of confusion stems from the barnardisation ruling. She identified two aspects to this; a conflict with the FoI legislation which exempts data controllers when a request involves additional processing effort, which barnardisation may be seen to do. Moreover, Gerz explained, the ruling further muddies the water about legally acceptable levels of anonymisation in the UK. Until (and unless) the matter is clarified by the House of Lords, she recommended erring on the side of caution when assessing disclosure risks for data that might be deemed personal. However data controllers

should also consider other exemptions from FoI; for example, where data is collected for ongoing research leading to publication, or where data may be withheld for public interest reasons.

Colin Pelton of NERC (Natural Environment Research Council) took up the theme, in a talk on the fit between FoI and **Environmental Information Regulations** (EIR). FoI specifically excludes environmental information which is covered by the Environmental Information Regulations 2004, enabling individuals and organisations to obtain environmental information held by public authorities. These regulations have clear relevance for NERC as the major funder of environmental research, training and knowledge transfer in UK universities and research centres. The EIRs have attracted far less publicity. Despite this they are in some respects more demanding for public authorities.

EIRs have a distinct history, being rooted in the Aarhus Convention and in turn a EC Directive on access to information, public participation and access to justice in environmental matters. These are very broadly defined to include opinion, advice, facts, measures, effects and analyses; relating to air, water, natural sites e.g. coastal or wetlands, flora and fauna (including crops, livestock, GMOs and biodiversity), the built environment, health, and emissions and discharges.

The UK regulations have much in common with FoI according to Pelton; both are overseen by the Information Commissioner (Information Commissioner's Office, n.d., Scottish Information Commissioner, n.d.), both provide a 20 working days response time, neither need be mentioned by requesters. Beyond that, EIR requests need not be in writing, EIR's have fewer exemptions ("exceptions" under EIR), and define "public authorities" more widely to include any person "with public responsibilities in relation to the environment". Furthermore, EIRs have no set charging regime, and cover all information produced or received by an authority, regardless of the reasons for its possession.

Pelton drew particular attention to the category of "emissions and discharges" information, since requests relating to that override exceptions that otherwise apply; for confidential information or internal proceedings, commercial or economic interests, or information normally withheld to protect the environment.

He concluded with some useful advice for requesters and authorities. Requesters would do better to mention EIRs rather than FoI when seeking environmental information because of their broader scope, even though requesters are not legally obliged to mention either. Authorities are best advised to deploy systems to manage business and scientific information effectively and transparently (email included) on an organisation-wide basis, and to track decision-making. Finally he recommended following rulings in this relatively new area of legislation.

Evidential Value and Authenticity

Michael Moss of HATII (Humanities Advanced Technology and Information Institute) at the University of Glasgow began the third session under the title "Beware the Smoking Gun - Was Old Mother Hubbard right?" He managed to breathe life into the phrase 'fiduciary protection' through the combination of fear and "wow" factor engendered by the use of emails as "smoking gun" evidence of public scandal; an example being the email correspondence used by New York Attorney General Elliot Spitzer as evidence in a case against Merrill Lynch leading to $1.4 billion in compensation and fines paid by brokerages and investment banks.

Moss's main theme was the need to manage risks properly, a key principle of digital curation. Spectacular failures in digital data management can, Moss argued, be partly placed at the feet of those who have neglected the analogue elements of data curation in pursuit of the digital; office procedures needed to be designed with other ideas in mind, such as "motivation" and "context management". Poor analysis of risk has led to simplistic analyses of systems and hence the loss of processes that have taken hundreds of years to evolve. This has been reflected in loss of context for much information; and the loss of the fiduciary protection once gained from "back office" working practices.

The responsibility for information governance is organisation-wide however, and is increasingly vested in audit and risk management committees. Information management is just a component of risk management, Moss stressed. Curation will be appropriate to the risk, and that risk assessed in light of the 'long arm of discovery' in global markets. That has led to an inclination to destroy any information sensitive to the risk of contingent liability; in other words fear of the "smoking guns" mentioned above can leave precious little data left to preserve in the proverbial cupboard. Rather than driven solely by fears, a balanced risk assessment should involve defining and asserting the business case for retaining information. He pointed to results of the Espida Project[4] as a useful resource for that.

The last workshop presentation was from **Burkhard Schafer**, of the Joseph Bell Centre for Forensic Statistics, Edinburgh University, on the prospects of '**modelling legal and archival knowledge in intelligent computer support tools'**. These he tied to the challenge of integrating the various kinds of expertise – legal and otherwise – on which day-to-day decision making processes may depend. One possible solution lies in computer-based decision-support tools that represent a theory of the relevant subject expertise. A conservator and archivist, again in theory, could then at any time access the relevant legal expertise necessary for him or her to make a decision.

However, research in legal Artificial Intelligence (AI) has shown that this model faces some difficulties, one of which is the different way lawyers, archivists, scientists or conservators frame a problem. Not only does such a system need to model the relevant legal expertise, it also needs the ability to "translate" the query by the archivist into the language of the law, with conceptual mismatches being a constant challenge.

Schafer drew inspiration from the novels of crime writer Jeffrey Deaver; on the importance of quantifying the significance of evidentiary traces by matching them with

[4] http://www.gla.ac.uk/espida/

databases of what they are traces of; and the message that pursuing one line of enquiry may preclude others. In curation, digitisation may change the characteristics of an object, removing DNA evidence for example. This implies that care needs to be taken in further curation; that further steps are minimally instrusive, and that all actions taken are recorded.

He related these principles to work on a small prototype system developed for the intersection between (criminal) law, conservation science and archival science, namely a searchable database intended to aid detection of the illegal sale of art and antiquities. This was designed to integrate, in the form of ontologies, the different conceptual schemes of police, legal, and arts professionals. It aimed to help relevant authorities to find objects that should "raise a red flag". The key question of the approach for Schafer is "does it scale?" given the range of expertise involved and the differences in law between jurisdictions. Moreover the output of such systems may not count as evidence in itself - a problematic matter where "reasonable suspicion" may be required to pursue further lines of enquiry.

There followed three breakout sessions where each of the three main sessions were discussed. In the session I attended Rena Gertz fielded queries on data protection and FoI and provided advice on some individual concerns about the difficult overlaps. Finally, **Mahesh Madhaven** of the JISC Legal service gave a helpful summing up of the day.

Conclusion

For those charting the rough sea between the data deluge and the dry ditch of technical obsolescence, this DCC workshop served as a reminder that data legislation can provide a way for data curators and managers to navigate through unseen dangers. At times that same legislation seems more like an iceberg of unfathomable mass than the gentle "landscape" of the workshop title, but events like this are a very useful opportunity to update one's navigation charts – and to hear tales of the sea monsters and hidden rocks to account for in a risk management plan.

References

Center for Research Libraries (CRL). (2007a). Certification of Digital Archives project. Retrieved November 26, 2007, from: http://www.crl.edu/content.asp?l1=13&l2=58&l3=142

Center for Research Libraries (CRL). (2007b). Trustworthy Repositories Audit & Certification: Criteria and Checklist; Retrieved November 26, 2007, from http://www.crl.edu/PDF/trac.pdf

DCC/DPE. (2007) Pilot audit programme. Retrieved November 26, 2007, from http://www.repositoryaudit.eu/participate/

The Guardian. (2007). "Privacy watchdog calls for power to carry out spot checks" by Patrick Wintour, Wednesday November 21, 2007. Retrieved November 26, 2007, from http://politics.guardian.co.uk/homeaffairs/story/0,,2214500,00.html

Hatcher (2007) Open Data. Retrieved November 26, 2007, from http://www.opencontentlawyer.com/open-data/

Information Commissioner's Office. (2007). Environmental Information Regulations information. Retrieved December 5, 2007, from http://www.ico.gov.uk/Home/what_we_cover/environmental_information_regulation.aspx

NESLi. (2007). Model licence. Retrieved November 26, 2007, from http://www.nesli2.ac.uk/ModelNESLi2LicenceMay07final.doc

Scottish Information Commissioner. (n.d.). Environmental Information Regulations information. Retrieved November 26, 2007, from http://www.itspublicknowledge.info/Law/EIRs/EIRs.asp