JLIS.it

The long-term preservation of the digital heritage: a case study of universities institutional repositories

Luciana Duranti

Digital preservation can be defined as the whole of the principles, policies, rules and strategies aimed at prolonging the existence of a digital object by maintaining it in a condition suitable for use, either in its original format or in a more persistent format, while protecting the object's identity and integrity, that is, its authenticity.[1] One might ask why authenticity is an issue for all kinds of digital objects. While it is obvious that authenticity is a necessary requirement in the preservation of records, the value of which as sources of evidence resides in their trustworthiness, why would it be for other types of digital heritage, such as publications, works of art, or games?

Although a digital entity that does not qualify as a record is not conceptually linked to the idea of trust, it still needs to have an identity that is certain and indisputable, and its manifestation must be stable, always equal to itself, intact. And here lies the problem, because, traditionally, these qualities reside in the original, the perfect entity, the first complete item to be issued, released or made public (be it a unique one or one of many), or in an authentic

[1] InterPARES Project Terminology Database, <http://www.interpares.org/ip2/ip2_terminology_db.cfm>.

JLIS.it. Vol. 1, n. 1 (Giugno/June 2010), p. 157–168.
DOI: 10.4403/jlis.it-12

copy of the original generated by a person with the authority to do so. However, as the concept of original disappears in the digital world, where can we look for the assurance that what we observe is what it claims to be?

We have no other choice than make inferences based on a variety of circumstances, but primarily on the integrity of the environment in which the digital entities in question reside and of the processes aimed to maintain them and to ensure the accountability of the person or organization responsible for them. This means that institutions must create mechanisms that allow for the determination of authenticity based on the trustworthiness of the source of the digital entities and the chosen method of their transmission through time, and then adopt the necessary strategies to preserve them in a sustainable way.

The selected preservation methodology must allow the preserved entities to continue to be readable and useable regardless of any technological changes to the underlying hardware or software environments, and the preserving organization to account for these changes, so that the entities may continue to be migrated to newer platforms as needed to avoid technological obsolescence.

To illustrate the issues digital repositories of cultural heritage will have to deal with, I am going to discuss an InterPARES 3[2] Case Study called "cIRcle." cIRcle is a digital repository for the management, dissemination and preservation of the intellectual output of a university and its community members.

A university institutional repository is defined as «a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the

[2]InterPARES 3 is the third phase of the InterPARES Research Project (1999-2012). While the first two aimed to develop theory and methods of authentic digital long term preservation, the third is testing the findings of the previous two in real situations. See http://www.interpares.org/ip3/ip3_index.cfm.

institution and its members » (Lynch 2003). Although their creation has been predicated in large part on the requirement of making available to the public research products that have been developed with the support of public money, and on the benefit to the university of showcasing its research, institutional repositories (hereinafter IR) have emerged in North America and Western Europe primarily because they are regarded by the university communities as a means of having access to products of scholarship and research, and as a locus for preserving such products and maintaining access to them over the long term (Lynch and Lippincott 2005; Westrienen and Lynch 2005). Therefore, in the past few years, the IRs have accumulated not only preprints and post-prints of articles, books, theses and dissertations, but also raw data files resulting from research, working papers, course syllabi, class notes, handouts, students' papers, committee meetings agendas and minutes, unpublished conference presentations and several other types of documentation that fall under the category of personal and university records, not only publications, and are preserved by the university archives and special collections. This mix of documentation and data creates severe challenges to the IR's continuing access and preservation (the very reasons why they exist) from several points of view.

The InterPARES research project has demonstrated that it is not possible to preserve digital materials, but only the ability to reproduce them. Reproduction involves different activities at different times in the life of the material. In the initial few years, it may consist simply of retrieving and reassembling the digital components that constitute the object to generate a copy or, if the object is technologically complex, as in the case of interactive and/or dynamic documents from the visual and performing arts and from the sciences, it may require its re-creation. However, when the digital format becomes obsolescent, it is necessary to either migrate the

digital object to a newer technology by changing its architectural structure or, in some cases, to emulate the behaviour of the old technology to access the object. Regardless, throughout the existence of the object, ongoing copying and transformative migration[3] are required for reasons of security (which is based on redundancy) and of continuing access. These activities raise several issues, among which the paramount ones are those of authenticity and intellectual rights. The authenticity of digital material is dependent upon the maintenance through time of its identity and of its integrity (Duranti 2005). The intellectual rights of the copyright owner are attached to the authentic version of the digital object and, specifically, to its documentary form, which is defined as the rules of representation that govern the expression of the ideas of the author in a stable and fixed manner (Duranti and Thibodeau 2006) .

Intellectual rights comprise several types of rights, but among them the ones that are affected by long-term preservation by means of constant transformative migration or emulation are the two major groups of intellectual rights: economic rights and moral rights. Economic rights are those that enable the copyright owner (not necessarily the author or creator) of a work to make commercial gain from the exploitation of that work (O'Hare 1982). Moral rights are those rights that the author or creator retains (regardless of whether the author still retains the economic rights) over the integrity of a work (**rights of reputation**)—such that no one, even the copyright owner, is allowed to distort, mutilate or otherwise modify the work

[3]Transformative migration is defined as "The process of converting or upgrading digital objects or systems to a newer generation of hardware and/or software computer technology" (InterPARES 2 Project, "Terminology Database: Glossary," available at http://www.interpares.org/ip2/ip2_terminology_db.cfm). The effects of transformative migration of the digital materials in an institutional repository are an important consideration insofar as any new additions or modifications to an existing work (even a work already in the public domain) can trigger new copyright considerations.

in a way that is prejudicial to the author's honour or reputation; the right to be associated with the work as its author by name or under a pseudonym and the right to remain anonymous (**rights of attribution**); and the right to refuse to allow the work to be used in association with a product, service, cause or institution in a way that is prejudicial to the author's honour or reputation (**rights of association**. (Rajan 2004)).

A recent census of college and university IRs in the United States has found that 70.8% of them do not have a policy for licensing content. In addition, there is no mention in the literature concerning IRs of the issue of authenticity through time, and none of them appears to have strategies in place for long-term preservation (Yakel *et al.* 2008). This is probably due to the fact that it is the belief of those who manage an IR that its content exists somewhere else, which is a safe presumption for preprints and post-prints, but certainly not for all those digital objects that are unique and often qualify as records, such as the official copy of theses, professors' and students' papers, etc.

Given the situation described above, it is necessary:

- to identify in which way digital preservation strategies as recommended by the major international research projects on the subject may infringe existing intellectual rights (economic and moral) legislation as it applies to published and unpublished material;
- to establish what long-term preservation measures would be possible in the context of the existing legislation and to test them on IRs in course of development to assess their impact on the continuing authenticity and accessibility of the digital material; and
- to determine what changes to the law are required to ensure

> that the proper long-term digital preservation strategies can be applied so that the research output of universities can remain attributable and accessible in its authentic form for as long as needed.

In order to do so, the InterPARES 3 project has selected as a case study an institutional repository called cIRcle, at the University of British Columbia (UBC).[4] As stated in the brochure publicizing it to the UBC faculty and students, cIRcle assembles various communities and collections. Communities are UBC departments, labs, research centres, schools and other administrative units. Within cIRcle, each community oversees one or more of its own collections, which contain items submitted to the IR. As currently envisioned, cIRcle's operational goal is to be able to accept, preserve indefinitely and provide continued readability and accessibility to virtually all published and unpublished digital objects created in any file format by or on behalf of the University, its faculty, staff or students—including preprints and post-prints of academic journal articles, other items such as theses, dissertations, departmental publications, technical reports, bulletins, conference proceedings, course notes and other learning objects, and raw research data. cIRcle has yet to develop, articulate and implement a maintenance plan that addresses this ambitious goal, and it has not attempted yet to address the issue of the protection of economic and moral rights in the context of long-term preservation. This situation makes of it the ideal candidate for the development of a preservation strategy for IRs that is sensitive to intellectual rights issues and for the testing of such a strategy.

As of November 6, 2009 there were 14,073 items in cIRcle totaling 130GB. This material is stored in DSpace, which is a database with a set of services to capture, store, index, maintain and make accessible a variety of entities in a digital format over the internet utilizing a

[4] https://circle.ubc.ca/.

controlled set of workflows and access permissions. Dspace is an open-source application, freely accessible at sourceforge.net, one of the largest open source software repositories on the net. It is written in Java, providing broad based support and compatibility with a broad base of internet browser; for a database back end it uses either Oracle – the industry leader – or Postgres, an open source relational database. The fact that DSpace is an open source application is good because of the authenticity issue.

It was stated earlier that the preservation of authenticity, to which all intellectual rights are linked, requires the protection of the identity and integrity of the material. Identity is not difficult to maintain overtime if the appropriate set of metadata is attached to the various digital entities and kept inextricably linked to them. Integrity is problematic not just to protect, but also to prove, because one has to rely on the integrity of the environment in which the entities reside. It is very hard to assess the integrity of an environment that is proprietary. In contrast, open source satisfies the legal requirements of objectivity, transparency, verifiability and repeatability for any process that is carried out in a digital environment (Carrier 2003; Ghirardini and Faggioli 2007; Kenneally 2001).

In DSpace, the records themselves are embedded in a hierarchical folder structure based on the collection. Contained within each folder is the original bitstream, a full text extract of the contents (used for searching), a thumbnail of the record for web presentation, and copyright information on the record. DSpace uses Preservation Services modules to verify the integrity of the stored files (utilizing a checksum to look for file corruption or alteration) and media filters to define file conversions. Records are accessed through the web via a persistent web address that allows researchers to link directly rather than having to use a database search every time. The bitstream format contains information on how the material in that

format is to be interpreted, allowing for control and granularity. For example, ".doc" may refer to more than one version of word, each of which presumably has different characteristics and functionality. Each bitstream format also has support level associated with it indicating how likely it is that the format will be accessible into the future given the toolset currently available to the system.[5] Digital forensics experts value open source, which, at the same time, allows modification and encourages dissemination, thereby making it possible to submit the software together with the digital entities presented as evidence, so that their accuracy can be tested promptly by anyone at any time. This is especially true when conversion or migration occurs, because it would allow a practical demonstration that the software could not simultaneously manipulate the content of the files while copying them and that nothing could be altered, lost, planted, or destroyed. Finally, open source is preferred because of the possibility of exchange of evidentiary material between the parties in the course of e-discovery.

Why should we care about issues of evidence and discovery? Because it is more than certain that, if an author feels that his or her intellectual rights have been infringed by the preservation measures taken by the institutional repository, he or she will want to see the issue solved in court. Undoubtedly, cIRcle procedures are in conflict with copyright legislation also regardless of preservation methods. This is because one has to consider that acting within copyright is different from policing copyright. Items are generally posted to

[5]The three levels are Supported, Known and Unsupported. Supported are those file types for which the institution has reasonable level of confidence to have the tools and/or techniques available to progress the files through future technology changes. Known formats are those that are recognized by the institution and in relation to which attempts are being made to create or obtain the tools necessary for future migration/access. Unknown file formats are those that will be preserved at the bitstream level only; it will be up to the researcher to obtain the software/tools necessary to view the files.

cIRcle by the author of a digital entity or the author's representative. Each submission requires the depositor to authenticate his/her authority to submit this work. cIRcle staff don't have enough time to verify copyright ownership for each item submitted, so cIRcle has to rely on the declaration of the depositor in order to operate. This is an act of faith, but it is a necessary one. Provided that cIRcle removes work upon notice of an infringement, and provided that cIRcle did not publish the infringing work knowingly, cIRcle should be protected from prosecution. Another issue has to do with materials that are scanned and uploaded to an IR. cIRcle's retrospective theses project involves the deposit of digitized theses and dissertations originally archived in print. It can be difficult to contact the authors of these items to obtain their permission to deposit. When authors cannot be contacted after due diligence in attempting to notify them, cIRcle may choose to proceed with publishing their work online. They do so assuming implicit permission by dint of the university's prevailing stewardship and provision of access to the item. Should the author request removal of the work from cIRcle after it has been published online, cIRcle remains obliged to honour the author's wishes and remove the work from its holdings.

But these are minor issues with respect to the problems presented for intellectual rights protection by the preservation strategies that cIRcle will have to adopt in the presence of a legislation that is still much behind the development of technology. cIRcle will have to begin separating the protection of the moral rights and that of the economic rights. To do so, its strategy will need to distinguish data integrity, which means that the content of the entities in the repository is not modified accidentally or intentionally during the regular maintenance and use activities, from duplication integrity, which means that the process of creating a duplicate of the data for preservation does not modify either intentionally or accidentally the form

and composition of the original entity. This reproduction would either be based on the principle of non-interference, which involves a non-transformative conversion, or on the principle of identifiable interference, which means that the method used does alter the entities but the changes are identifiable (Casey 2002). The application of both principles, by ensuring the creation of authentic copies, would allow for the protection of moral rights, which cannot be renounced. As it regards economic rights, any preservation activity would infringe the law as it stands now, thus, the only solution at this time is to obtain the permission of the copyright owners. The InterPARES Project has made a submission to the Commission of Industry Canada and the Department of Canadian Heritage responsible for updating the Canadian copyright act, requesting that specific attention be given to the problems presented by the long term preservation of authentic digital entities.[6] In the meanwhile, it is conducting an inventory of all the items in cIRcle to identify their nature and characteristics, content, current licence, attached digital rights management, etc. in order to develop an intellectual property policy and a preservation plan consistent with it. The research conducted on cIRcle and its results will be accessible on a dedicate web site named "University Institutional Repositories: Copyright and Long Term Preservation", accessible at http://uir-preservation.org/.[7]

References

AUSTIN, G. W. (2005), "The Berne Convention as a Canon of Construction: Moral Rights after Dastar", *NYU Annual Survey of American Law*, 61, pp. 101–139.

BURKITT, D. (2001), "Copyrighting Culture – The History and Cultural Specificity of the Western Model of Copyright", *Intellectual Property Quarterly*, 2, pp. 146–186.

[6]Copyright Consultation statement, http://copyright.econsultation.ca .

[7]For more information on copyright and moral rights as related to preservations see: (Austin 2005; Burkitt 2001; Christie 1995; Davies 2002; Dreier 1998)

CARRIER, BRIAN (2003), *Open Source Digital Forensics Tool. The Legal Argument*, p. 7, http://www.digital-evidence.org/papers/opensrc_legal.pdf.

CASEY, E. (2002), "Error, uncertainty and loss in digital evidence", *International Journal of Digital Evidence*, 1, 2.

CHRISTIE, A. (1995), "Reconceptualising Copyright in the Digital Era", *European Intellectual Property Review*, 17, 11, pp. 522–530.

DAVIES, G. (2002), *Copyright and the Public Interest*, London: SweetMaxwell.

DREIER, T. K. (1998), "Adjustment of Copyright Law to the Requirements of the Information Society", *IIC*, 29, 6, pp. 623–639.

DURANTI, LUCIANA (ed.) (2005), *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project*, San Miniato, Italy: Archilab.

DURANTI, LUCIANA and KENNETH THIBODEAU (2006), "The Concept of Record in Interactive, Experiential and Dynamic Environments: the View of InterPARES", *Archival Science*, 1, pp. 13–68.

GHIRARDINI, ANDREA and GABRIELE FAGGIOLI (2007), *Computer Forensics*, Milano: Apogeo, p. 230.

KENNEALLY, ERIN (2001), "Gatekeeping Out Of The Box: Open Source Software As A Mechanism To Assess Reliability For Digital Evidence", *Virginia Journal of Law and Technology*, 13, 6, pp. 34–35, http://www.vjolt.net/vol6/issue3/v6i3-a13-Kenneally.html.

LYNCH, CLIFFORD A. (2003), "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age", *ARL Bimonthly Report*, 226, pp. 1–7, http://www.arl.org/bin~doc/br226.pdf.

LYNCH, CLIFFORD A. and JOAN K. LIPPINCOTT (2005), "Institutional Repository Deployment in the Unites States as of Early 2005", *D-Lib Magazine*, 11, http://www.dlib.org/dlib/september05/lynch/09lynch.html.

O'HARE, MICHAEL (1982), "Copyright and the Protection of Economic Rights", *Journal of Cultural Economics*, 6, 1, pp. 33–48.

RAJAN, MIRA SUNDARA (2004), "Moral Rights in Information Technology: A New Kind of 'Personal Right'?", *International Journal of Law & Information Technology*, 12, 1, pp. 32–54.

WESTRIENEN, GERARD VAN and CLIFFORD A. LYNCH (2005), "Academic Institutional Repositories: Deployment Status in 13 Nations as of Mid 2005", *D-Lib Magazine*, 11, http://www.dlib.org/dlib/september05/westrienen/09westrienen.html.

YAKEL, ELIZABETH *et al.* (2008), "Institutional Repositories and the Institutional Repository: College and University Archives and Special Collections in an Era of Change", *The American Archivist*, 71, 2, p. 344.

About

The author

Luciana Duranti
The University of British Columbia. School of Library, Archival and Information Studies (SLAIS)
Email: luciana@interchange.ubc.ca

Web: http://www.lucianaduranti.ca/

The paper

Date submitted: 2010-03-06

Date accepted: 2010-05-03

Links checked: 2010-05-22

Date published: 2010-06-15