[Michael A. Fraser, “Identifying Digital Objects” in Dutton, W. H., and Jeffreys, P. W. (2010 forthcoming) (eds), *World Wide Research: Reshaping the Sciences and Humanities*. Cambridge, MA: The MIT Press.]

Identifying Digital Objects

Michael A. Fraser

Introduction: Defining and Managing Digital Objects

Establishing the policies, infrastructure and tools to create long-term persistent, accessible and reusable digital data is one of the great challenges for the research and library communities today. This challenge applies to the petabytes of data generated from large-scale scientific instrumentation as much as it does to individual journal articles, and is as relevant to the humanities and social sciences as it is to the physical and bio-sciences.

In this essay, a digital object is defined as a combination of one or more files with metadata (including identifiers) which can be stored in - and retrieved from - a file system (Lagoze, Payette, Shin, et al. 2005; Kahn and Wilensky 2006). However, any concise definition risks hiding the complexity inherent to digital objects.[1]

The *raison d’etre* for a digital repository is the management of digital objects, and most repository systems have well-defined processes for ingesting, storing, exposing and accessing them objects. Other systems, whether managed within an institution or by third-party ‘community’ services, are increasingly providing *de facto* repository services. For example, a virtual learning environment (see chapter 10) is unlikely to have been designed as a digital repository, in any sense of the term as understood

1 This essay follows the digital object approach of the Functional Requirements for Bibliographic Records (FRBR) in making a distinction between the abstract 'work' and one or more 'expressions' of work manifested digitally, see IFLA (2007); Renear and Dubin, 2003; Bide et al., 2006.

within the library and information science community – but at least it is a managed system.

The multitude of disks in offices, laboratories, computer rooms, and dangling from key rings also contain myriad tangled digital objects, clearly not managed within a repository system.

For the majority of digital objects created in the course of employment, an institution clearly has a vested interest in their management. A proportion of these objects, probably a small proportion overall, will find their way in some version or another to formal digital repository systems; but most will not. It is imperative that the relevant parties (researchers, information scientists, institutional managers) collaborate on supporting the lifecycle of an institution's digital asset. This support must start when a device (desktop, mobile phone, camera, sensor, scientific instrument, etc.) brings the object into being, rather than only when the object is offered to a digital repository in a one-off event (Hockx-Yu 2006).

Assurance through Digital Object Identification and Authentication

At the heart of the challenge in creating persistent, accessible and reusable data lies the identification of digital objects. This identification is fundamentally about assurance (Renear and Dubin 2003) – a match between the data and a set of assertions about those data – and is essential for:

- *Discovery and access*. People and systems need to be able to discover and retrieve digital objects within a networked environment to match pre-defined criteria;
- *Provenance and assurance*. Levels of assurance are required to determine what the digital object is, from where it originates, how it has evolved, and the

relationships beyond itself;

- *Reusability*. Digital objects are increasingly reused and re-purposed, with new objects derived from aggregation, enhancement, or subtraction
- *Preservation*. The unpredictability of when any of the previous elements might be desirable and the increasing complexity of the process of identification as software, data formats and hardware technologies advance must be dealt with.

The identity of a digital object comprises attributes related to its: name (identifier); intellectual content; format(s) in which it is represented; and internal and external relationships, including how the object might be processed, and any earlier or later versions, revisions, and derivatives.

The authentication of digital objects is problematic. There are various robust schemes for authenticating people within the digital sphere (the most secure of which depend on authentication outside of the digital world – for example, using some combination of birth certificate, passport, and affiliated institution). There is no requirement, however, to register the birth of a digital object; no universally acknowledged passport and visa system that permits a digital object to move from one domain to another. Moreover, existing systems for the authentication of digital objects have been bound up with digital rights management, placing greater emphasis on the licence or authorisation that sits between the user and the object.

Examples of Digital Object Identifiers

The most basic identifier is the file name. However, many file-naming conventions are arbitrary, specific to a system or application, or based on parochial practice; have no defined process for change; tend not to make explicit any relationships with other

digital artefacts; and are, therefore, hardly mobile once the object leaves its immediate local context. There have been various attempts to create persistent naming schemes for digital objects. One of the best known is the digital object identifier (DOI),[2] comprising an address, resolver, and set of policies governed by an international foundation (Vitiello 2004). The business model for maintaining the DOI system is aimed at publishers – for example with DOIs often being assigned to digital journal articles which are a reasonably simple type of digital object. The granularity at which persistent identifiers are applied is an issue for composite digital objects. The relationship between identifiers must maintain the integrity of a digital object in its immediate context (Arnab and Hutchinson 2006).

Establishing the Provenance of Digital Objects

In the study and trade of art works and antiquities, provenance is the process by which the history, context and ownership of an object is established and its realisation is, in effect, a very rich set of metadata. The issues are very similar with respect to the identification of digital objects within the scholarly environment.

The provenance of a digital object is required to establish the process by which the object and its constituent parts were created (e.g. workflows, subsequent additions or subtractions), as well as its ownership – which can be complex when data arise from large-scale consortium projects. Documenting the provenance of a digital object must not only capture the most useful information for establishing identity and trust, but also ensure that the often extensive provenance metadata, or a reference to them, are carried with the digital object.[3] There must also be a trusted means by

2 See <http://www.doi.org/>.

3 Software applications increasingly embed metadata within files. However, open standards and tools for managing embedded metadata are also required.

which provenance data are created or added, such as digital signatures (Arnab and Hutchison 2006).

Future Bridges between the Digital and Physical

Although this essay has focused on the attributes necessary for the persistent identifying of digital objects, similar digital attributes assist in identifying non-digital objects. In the world of increasingly pervasive computing, various bridges exist between digital and physical objects. For instance, the combination of radio frequency identity electronic tags, microsensors, embedded networked systems, global positioning systems, and corresponding receivers blurs the boundaries between the physical and the digital realms. Location and context-aware relationships between such objects can be defined and recorded for history. Bruce Sterling, (2004) has coined the term *spime* to describe objects that can be "precisely located in space and time. They have histories. They are recorded, tracked, inventoried, and always associated with a story".

The resolution of issues around the identification of digital objects has barely begun. Imagine, however, the emergence of an "Internet of Things" (ITU 2005) composed of anything from books, to clothing, to flora and fauna, each streaming metadata about what it is, whom it's with, and where it's been. Welcome to the metadata deluge.

References

Arnab, Alapan and Andrew Hutchison (2006) “Verifiable digital object identity system”. *Proceedings of the ACM workshop on Digital rights management 2006, Alexandria, Virginia, October 30 - 30, 2006*: 19-26. <http://doi.acm.org/10.1145/1179509.1179514>.

Bide, Mark, Michael Fraser, Deborah Kahn, Hugh Look, Howard Noble, Sally Rumsey and Frances Shipsey (2006). *Scoping Study on Repository Version Identification (RIVER): Final Report*. JISC Working Group on Scholarly Communication. Bristol: Joint Information Systems Committee. <http://www.jisc.ac.uk/uploaded_documents/RIVER%20Final%20Report.pdf>.

Hockx-Yu, Helen (2006). “Digital preservation in the context of institutional

repositories". *Program: electronic library and information systems* 40(3): 232 - 243.

IFLA (2007 [1997]) *Functional Requirements for Bibliographic Records: Final Report*. The Hague: International Federation of Library Associations and Institutions (IFLA). <http://www.ifla.org/VII/s13/frbr/>.

ITU (2005) *The Internet of Things, ITU Internet Report*. Geneva: International Telecommunication Union. <http://www.itu.int/publ/S-POL-IR.IT-2005/e>.

Kahn, Robert and Robert Wilensky (2006). "A framework for distributed digital object services". *International Journal on Digital Libraries* 6(2): 115-123.

Lagoze, Carl, Sandy Payette, Edwin Shin, and Chris Wilper (2005). "Fedora: an architecture for complex objects and relationships". Draft of submission to *Journal of Digital Libraries* Special Issue on Complex Objects, version 6, 23 August 2005. <http://arxiv.org/abs/cs/0501012v6>.

Lynch, Clifford A. (2001). "When documents deceive: trust and provenance as new factors for information retrieval in a tangled web". *Journal of the American Society for Information Science and Technology* 52(1): 12-17.

Renear, Allen and David Dubin (2003). "Towards identity conditions for digital documents" In S. Sutton (ed), *Proceedings of the 2003 Dublin Core Conference, Seattle, WA, October 2003*. University of Washington. <http://www.siderean.com/dc2003/503_Paper71.pdf>.

Sterling, Bruce (2004). "When Blobjects Rule the Earth". SIGGRAPH, Los Angeles, August 2004. <http://www.viridiandesign.org/notes/401-450/00422_the_spime.html>.

Vitiello, Giuseppe (2004). "Identifiers and identification systems: an informational look at policies and roles from a library perspective. *D-Lib Magazine* 10 (1). <http://www.dlib.org/dlib/january04/vitiello/01vitiello.html>.

About the author

Michael A. Fraser is Head of Infrastructure Systems and Services at Oxford University Computing Services.