Preserving Change: Observations on Weblog Preservation

Yunhyong Kim
Humanities Advanced Technology and Information Institute
University of Glasgow
Glasgow, UK
Yunhyong.Kim@glasgow.ac.uk

Seamus Ross
Humanities Advanced Technology and Information Institute
University of Glasgow, Glasgow, UK
&
Faculty of Information
University of Toronto
Toronto, Ontario, Canada
seamus.ross@utoronto.ca

ABSTRACT
In this article, we revisit concepts introduced within the digital preservation literature, such as Open Archival Information System (OAIS) reference model, and Preservation Metadata Implementation Strategy (PREMIS), to examine their continued applicability to the preservation of dynamic web content such as weblogs.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – Standards.

General Terms
Management, Design, Human Factors, Standardization, Theory.

Keywords
digital preservation, digital curation, designated community, authenticity, intellectual entity, archive, web archive, blog, weblog

INTRODUCTION
Current preservation approaches tend to be largely data object oriented, relying on the notion that data can be reasonably reduced to a manageable discrete set of objects accompanied by formal syntactic, semantic and pragmatic attributes that constitute the original object's content and characteristics necessary for validating authenticity, managing rights, and enabling access and use (e.g. see [1], [6]). Now, the dynamic web environment (e.g. blogs, wiki, networking platforms) enables us to capture data objects at finer levels of communicative granularity. Continuing to capture each of these bits as a discrete entity/object imposes independent object identities on pieces of information that, in the past, would have only been considered to have meaning as part of the whole intellectual process. It may be time to re-examine the established approaches to determine whether they are still valid in the context of web archiving initiatives (e.g. the Minerva Project[1], Internet Archive[2], UK Web Archive[3], Arcomem[4], BlogForever[5], Memento Project[6], LiWA[7]) that have been increasingly trying to create solutions for social media archival situations.

iPRES2011, Nov. 1–4, 2011, Singapore.

OAIS: A BRIEF SUMMARY
The Reference Model for an Open Archival Information System (OAIS) [1] is a conceptual model for a preservation-aware archival system developed by the Consultative Committee for Space Data Systems (CCSDS) (accepted as an ISO standard in 2003[8]). It has been adopted by several well-known preservation projects in recent years (e.g. CASPAR[9], SHAMAN[10], SHERPA DP2[11] and the Planets Interoperability Framework [9]). To be compliant to the model (see [1]), "the OAIS must: 1) negotiate for and accept appropriate information from information producers; 2) obtain sufficient control of the information needed to ensure long-term preservation; 3) determine which communities should become the Designated Community and, therefore, should be able to understand the information provided; 4) ensure that the information to be preserved is independently understandable to the Designated Community; 5) follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated copies of the original, or as traceable to the original; and, 6) make the preserved information available to the Designated Community."[12].

PREMIS DATA MODEL
The PREMIS (Preservation Metadata: Implementation Strategies) working group was sponsored by OCLC Online Computer Library

[1] http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html

[2] http://www.archive.org

[3] http://www.webarchive.org.uk

[4] http://www.arcomem.eu

[5] http://www.blogforever.eu

[6] http://www.mementoweb.org/

[7] http://www.liwa-project.eu/

[8] ISO/DIS 14721

[9] http://www.casparpreserves.eu

[10] http://shaman-ip.eu/shaman/

[11] http://www.sherpadp.org.uk/sherpadp2.html

[12] This content from [1] has been condensed to save space.

Center and Research Libraries Group (RLG), to develop a core set of preservation metadata applicable to a wide range of digital preservation contexts. The resulting standard [6] was intended to comply with the OAIS model (Section 2.1), while targeting metadata that capture preservation processes, such as the preservation level associated with an object. While descriptive and technical metadata are also key concepts in the standard, PREMIS recommends the use of previous standards to meet requirements for these, focusing on preservation levels and processes, rights, and object properties and relations to be preserved. A large amount of the effort in PREMIS remains with object modeling. While notions of agents, events and rights are discussed within the standard, detailed information is not provided. The model relies on the concept of an **intellectual entity** as a single intellectual unit to be managed within the archive.

OBSERVATIONS ON WEB ARCHIVING

While many web archives have claimed compliance with the OAIS model (Section 2.1), this can be accepted only on the most generous terms: 1) while access can be blocked, there is almost never any explicit negotiation between information producers and existing web archives: the information is obtained through procedures for "copying the website" [7]; 2) the lack of negotiation means that the archive's rights to manipulate harvested pages for preservation purposes becomes ambiguous, and introduces an unpredictable gap between the archive's "authentic copy" and the material at the time of creation; 3) by equating inaction of creators with permission for the archive to retain the material, the integrity of the archive's content is put at risk, as any material (e.g. an image within a blog post) may be later requested to be removed; 4) the selection of a designated community is also largely washed over in the web context: in the case of blogs, there is no clear long-term readership, as evidenced by the constantly fluctuating statistics available through search services such as Technorati[13]; 5) the long term deterioration of integrity (through missing objects and lapsed URLs) will result in semantic gaps in the knowledge base; 6) the notion of an intellectual entity is also blurred (e.g. see [2]): new blog posts are added to blogs periodically, previously submitted posts and comments are modified, deleted, and rearranged, changing rights, content and semantics.

As a solution for point 6), some have introduced the notion of archiving versions at varying times as independent intellectual entities. Others have tried to break down the blog into smaller intellectual entities (e.g. posts, comments, embedded objects). This approach could lead to: 1) an unmanageable increase in data storage, 2) many instances of semantically incomplete objects (posts often make sense only in the context of other posts, and even more so for comments and embedded images), and, 3) millions of objects with minor differences between them.

TOWARD PRESERVING CHANGE

We emphasise the *predominance of change* as a core characteristic of today's digital information environment. Change has, of course, always been an integral part of digital information. As we access, save, and transmit information, we cause change and deterioration. To ensure that information does not change from its original state has become essentially impossible [5]. The core purpose of preservation is, not to capture the illusory static steps in between changes, but to ensure that we capture the change itself, and, preserve how changes might propagate other changes. The time dimension in the preservation of the webpages has already been recognised[14,15] but the current paradigm is to understand change as the time-stamped objects in selected states. Ontologies have been proposed to capture events and relations between objects (e.g. see Event Ontology[16] used to represent musical performance; ABC Ontology proposed for preservation [3]). While ontologies provide a step in the right direction, they still describe transitions of object states. Our contention is that objects are symptoms of dynamic processes generated by the medium through which they are broadcast. These need to be captured as recurring patterns within medium dependent event windows that go beyond object boundaries (Figure 3.1).

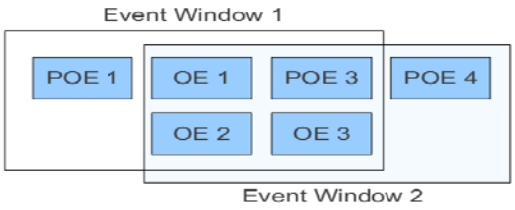

Figure 3.1. Windows of size 3 surrounding physical object events (POE) and other events (OE).

To quote Marshall McLuhan: "the medium is the message" [4]. There is semantics beyond the content of a message: the emergence of so many different channels of communication (e.g. blogs, twitter and facebook) may be a testament to the part that the medium plays in conveying meaning and purpose.

ACKNOWLEDGMENTS

The research leading to the discussion in this paper was conducted as part of the BlogForever project funded by the European Union's Seventh Framework Programme (FP7-ICT-2009-6) under grant agreement n° 269963.

REFERENCES

[1] CCSDS (2002) "Reference Model for an Open Archival Information System (OAIS)", *CCSDS 650.0-B-1* (2002): http://public.ccsds.org/publications/archive/650x0b1.pdf

[2] Hank, C., Choemprayong. S., and Sheble, L. (2007) "Blogger perceptions on digital preservation." In Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (JCDL '07). ACM, New York, NY, USA. http://doi.acm.org/10.1145/1255175.1255276

[3] Lagoze, C. and Hunter, J. (2002) "The ABC Ontology and Model", Journal of Digital Information, Vol 2, No 2, http://journals.tdl.org/jodi/article/viewArticle/44

[4] McLuhan, M. (1964) Understanding Media. Routledge, London.

[5] Montague, L., Nicchiarelli, E., Matthezing, H., Kummer, R., Puhl, J., & Roberts, B. (2010b) "The concept of significant

[13] http://technorati.com/

[14] Compare with approaches at http://www.mementoweb.org/

[15] Denev et al. (2011) The SHARC framework for data quality in web archiving. *VLDB Journal*, 20(2):183—207.

[16] http://motools.sourceforge.net/event/event.html

properties", The National Archives, UK & The Austrian National Library

[6] PREMIS Editorial Committee (2011) PREMIS Data Dictionary for Preservation Metadata Version 2.1, *PREMIS Editorial Committee*: http://www.loc.gov/standards/premis/v2/premis-2-1.pdf

[7] Roche, Xavier (2006) "Copying Websites", *Web Archiving*, Julien Masanes (ed.) Database Management and Information Retrieval, Springer, Pp 93-114: http://www.springer.com/computer/database+management+%26+information+retrieval/book/978-3-540-23338-1

[8] Wilson, Carl (2008) "Planets Interoperability Framework Guidelines for Service Wrapping", Planets Project, England: http://sherpa.bl.uk/113/01/Planets_IF6-D2_GuidelinesForServiceWrapping_Ext.pdf