Collecting Usage Data for Digital Preservation

Muriel Foulonneau
Tudor Research Centre
29, av John F. Kennedy
L-1855 Luxembourg
Luxembourg
+352 42 59 91 - 1
muriel.foulonneau@tudor.lu

ABSTRACT

While IT environments are moving towards personalized and context-aware adaptive content and services, digital preservation systems should go beyond the current mechanisms to preserve digital objects. Social and personal experiences need to be investigated as part of the context of digital resources, i.e., the way in which a resource was used and perceived, by retaining usage data for instance. Overall, users have to be further involved in the digital preservation processes, in the creation of context metadata, in the storage and migration of resources, in particular for personal archives.

Categories and Subject Descriptors

H.4 [**Information Systems**]: Information Systems Applications – *miscellaneous,* H.5.1 Multimedia Information Systems

General Terms

Keywords

Digital preservation, content centric networks, e-assessment, usage data, digital archiving

1. INTRODUCTION

Digital preservation has become a major issue for different types of users, researchers, individuals willing to preserve their personal archives, and companies with concerns related to litigation and knowledge management for instance. At the same time, IT environments are moving towards personalized and context-aware adaptive content and services. The nature and processes associated to digital preservation may have to evolve accordingly, to further involve users in the preservation processes.

2. PERSONALIZED AND CONTEXT AWARE DIGITAL OBJECTS

Digital services are more and more emphasizing users and their experience with digital content. They are changing substantially the content representation, the way in which it will be transferred and consumed in the future. This will modify completely our sense of what digital objects are. Digital objects will be adaptive, recomposed (content mashups), replicated across networks (content centric networks, see Jacobson et al., 2009, content objects, Zahariadis et al., 2010). The experience of users will not be directly in line with the bitstreams, which can be stored and preserved because context is driving the representation and/or dynamic composition of digital objects. At the same time, some researchers and professionals have been working on the emotional components of digital objects and archives. This relates to the selection and appraisal of resources, but also to the preservation of the perception of content rather than the bitstreams, which produced the emotions (Van den Broek et al., 2010). The preservation would then focus on the social impact and personal experiences of content.

This involves user tagging, eye tracking, log records and applying any type of technique, that can record usage activities and the perception of digital objects by one (personal perception and experience) or more persons (social perception and experience).

Personalization systems currently use access logs or sales tracking to identify usage patterns and recommend content or a particular representation of content. For instance, a book was a best seller, whereas only a few copies of another book were sold. This makes both books significantly different resources, because their social impact was different. All the same, if one could take pictures of the faces of visitors of Le Louvre in front of the Mona Lisa painting, this may tell a lot on the painting and the social experience it created. This strategy was used to create the trailer of the very successful movie “Paranormal activity” (2007). In addition to selected extracts of the movie, the authors of the trailer recorded the audience of a theatre where the movie was played.

3. USAGE DATA: THE CASE OF E-ASSESSMENT RESOURCES

A concrete example of usage data archiving can be found in the domain of computer-based assessment or e-assessment. Assessment items, especially in high-stake assessment (e.g., TOEFL) are usually calibrated, i.e., they are tested by sample users to determine their difficulty and discrimination potential for instance. But for security reasons (cheating risk), items should not be given too many times. They should also not be given to candidates, who may have been in contact with previous test candidates. The conditions (context) under which tests were carried out may also have allowed candidates to copy and share test questions. The fact that an item was used in a test, the test context, and information related to the test candidates could therefore be stored and archived (e.g., for future proof, reuse, or longitudinal studies). The IMS-QTI metadata format already includes usage data associated to assessment resources, in additional to more traditional descriptive metadata (see for instance Sarre et al., 2010).

The results of modern test items (including interactive test items or serious games for instance) also depend on candidates' behavior, which can be captured by logs, eye tracking, etc. All the user tracks become endless quantities of data, especially if the usage context is included.

This calls for defining the digital objects from their usage track record and redefining preservation the other way around, from the social impact and personal experience of objects. Users become one of the sources of value for digital resources.

4. USER PARTICIPATION IN THE DIGITAL PRESERVATION PROCESSES?

Users can therefore contribute to the preservation process in a multiplicity of ways, before ingest in a digital preservation system and even over the course of the digital preservation process.

As described above, in order to preserve the perception and/or social existence of resources, it is necessary to capture users' activities and / or perception. Many applications aim to capitalize on the "wisdom of the crowd" to annotate content, through the correction of OCR for instance or metadata games, such as Ontogames (Siorpaes et al., 2008).

Moreover, users can be involved in the preservation process. They store content on their own devices. This multiplies the number of existing copies of digital resources. The replication of resources on many devices can support digital archiving, migration, quality control, and the recovery of content over time, provided that content identification and provenance is sufficiently documented inside digital resources.

However, user participation mechanisms have to be designed in order to limit associated risks (e.g., on authenticity and accuracy). Moreover, according to the digital preservation constraints (policy and rules), it may be more or less relevant or focused on one specific task.

In the domain of personal archiving (Marshall et al., 2006), users are involved in the creation of contextual data for instance. The emotional value of resources can be considered as important information to contextualize them. It raises questions related to the techniques to capture it, user acceptance, and rendering for instance.

Private companies however have extremely different constraints and objectives. Major challenges of private companies relate to trust, security and business value of preservation (e.g. Colet, 2010). Although the constraints are quite different, the way in which resources were used (e.g. by computer programs to render a specific representation of resources) is still relevant.

5. CONCLUSIONS

All these evolutions raise research questions, in particular related to the value of digital objects in the future, the selection, representation, and interpretation of usage data, finally, the involvement of users at different stages of the digital preservation process, together with the associated risks (e.g. on authenticity or accuracy). Tomorrow, nobody will ever consume the same representation of the resource preserved, but rather a specific adaptation, potentially unique, to the user personality and current context. Complex mechanisms will be embedded in "content", which will instead be able to react to its environment, just like software or agents. Digital preservation should then focus on preserving software, services and user experiences.

6. REFERENCES

[1] Colet, L. 2010. Electronic Records Management in Luxembourg: Challenges and Perspectives. ERCIM News 2010(80)

[2] Jacobson, V., Smetters, D.K., Thornton, J.D., Plass, M.F., Briggs, N.H. and Braynard, R.L. 2009.Networking named content. In Proceedings of the 5th international conference on Emerging networking experiments and technologies, pp 1-12.

[3] Marshall, C. C., S., and Brun-Cottan, F. 2006. The Long Term Fate of Our Digital Belongings: Toward a Service Model for Personal Archives. In Proceedings of IS&T Archiving 2006, (Ottawa, Canada, May 23-26, 2006), Society for Imaging Science and Technology, Springfield, VA, 2006, pp. 25-30.

[4] Sarre, S., and Foulonneau, M. 2010. Reusability in e-assessment: Towards a multifaceted approach for managing metadata of e-assessment resources. In Proceedings of the Fifth International Conference on Internet and Web Applications and Services, ICIW 2010, May 9 - 15, 2010 - Barcelona, Spain.

[5] Siorpaes, K., and Hepp, M. 2008. OntoGame: weaving the semantic web by online games. In The Semantic Web: Research and Applications, Springer, 2008, pp. 751-766.

[6] Van den Broek, E., van der Sluis, F., and Schouten, T.E. 2010. User-Centered Digital Preservation of Multimedia. ERCIM News 2010(80).

[7] Zahariadis, T., Daras, P., Bouwen, J., Niebert, N., Griffin, D., Alvarez, F., and Camarillo, G. 2010. Towards a Content-Centric Internet. In Towards the Future Internet, G. Tselentis et al. (Eds.) IOS Press.