Report of the PERSIST Workshop 'Demand driven selection for long term preservation of born digital cultural heritage',

Granada, DigitalHeritage2015 Congress, 29 September 2015

1. Introduction

First it should be stated that PERSIST was mentioned quite prominently in the keynote to the DH2015 conference on 28th of September by Guy Berthiaume, Director of the Library and Archives of Canada. He stated that PERSIST is a very important initiative and that the LAC supports the project wholeheartedly. This set the scene quite nicely for our workshop.

The workshop itself took place on day 2 of the DH2015 Congress and was moderated by Marco de Niet (DEN) and Clément Oury (ISSN IC), with preparatory support from Vincent Wintermans (Unesco NL) and Julia Brungs (IFLA). The duration of the workshop was 3 hours, and there were about 20 participants. The two groups that were represented most were heritage researchers (e.g. archaeologists) and professionals working in memory institutions (the archives, libraries and museums domains were all represented).

The workshop opened with a general introduction to PERSIST and the work of the Content Taskforce so far, followed by a general discussion about the way the participants use digital collections/resources in their daily work. From that perspective of their own use, they were invited to help assess criteria for to selection of born digital heritage for long term archival purposes.

2. General comments

During the first part of the workshop, some general comments were made about the intention to publish guidelines for selection in the context of PERSIST. Some participants questioned the usefulness of such generic guidelines that would try to address institutions across the globe. Wouldn’t they become too generic to be really useful in the specific, practical context of an institution? In the guidelines, we should explain carefully what the scope and ambition of the guidelines are. We’ll notably need to explain that these general guidelines should act as a reference from which more precise guidelines (e.g. dedicated to a specific country/institution type) may be drawn; and against which already existing guidelines may be assessed.

Another remark addressed the concept of selection. Some participants questioned the usefulness of selection as the way forward. Is it really (conceptually) possible to make useful selections for archiving purposes based on assessment of value for the future? Digital resources do not become cultural heritage because memory institutions put a stamp on them. Two approaches were discussed at the workshop how to address this in the guidelines:

- selection cannot take place without agreements with the creators of the digital objects that are being preserved. Memory institutions need to be more active 'upstream' where the data is created. This works in two ways. On the one hand, creators should be made more aware of what it takes to preserve digital data over time (e.g. use of specific standards). On the
other hand, memory institutions should have an open mind for the needs of creators of data to keep their data alive (some archaeologists at the workshop complained that no memory institution in their country has provided an opening to discuss conditions and facilities to help preserve the data they produce during their excavations). Either way, the selection of what will be preserved will then be done by the creators, not by the institutions that archive the resources.

- another way of looking at this, is that cultural heritage is based on assessment by the public over time. From that perspective, selection cannot be done beforehand and what we need is good querying options afterwards in large undefined sets of archived digital resources (cf. the approaches to archiving the web and social media).

Despite these remarks about the role of memory institutions in selecting, it was agreed that in the physical world, selection for future re-use has been a core expertise of memory institutions, and it is useful to have this discussion about selection by memory institutions to address the usefulness and development of this expertise into the digital realm. There are two pitfalls here: memory institutions can on the one hand be focussed too much on the general principles and not do anything practical; on the other hand, they may focus on the selection of some really specific digital resources and lose sight of the wider context. From this perspective, it makes sense to consider heritage to be the outcome of the process of careful heritage assessment in the present. It should be noted however that although this process will/ought to be initiated by professionals, other parties (incl. the users) should be included.

Another remark that relates to 'value assessment' concerned the use of the word 'value'. Value has a strong ethical dimension, but this dimension is not yet properly addressed in the Guidelines. One of the ethical issues that was mentioned during the workshop was that selection can be abused for political reasons. Selection may not be neutral. That is why it is important to stress that archiving cannot be the sole responsibility of a single institution and that the wider context (the example of fake websites created by political opponents was given during the workshop) needs to be addressed and incorporated in archiving activities. All in all, the ethics of value assessment and selection that result from it need to be addressed in the guidelines more prominently.

Besides, it was discussed what the position of museums is in the context of the PERSIST Guidelines, and what their perspective on long term archiving is. This topic seems more appropriate for archives and libraries than museums. However, it was agreed that esp. in the area of born digital art, libraries and archives cannot, in the current context, provide a solution for archiving this type of cultural heritage. It would be good to describe the position of museums in relation to digital archiving in a more explicit way.

A last comment related to the question of the institution’s mandate. If the selection of a content should be decided considering the institution’s responsibility, there is a risk that some content outside any organisation’s scope – for cultural or political reasons – may be forgotten. Example was given of resources documenting the apartheid, that South-African institutions, during the apartheid period, were not supposed to archive. Therefore, the PERSIST project should also question and provide recommendations on institutional mandates – but it is more the role of the PERSIST Policy taskforce.
3. Specific recommendations for editing the draft guidelines

The guidelines address the similarities and differences between the domains (archives/libraries/museums), but this provides only a partial view on selection issues. There are other boundaries that are both relevant and under pressure because of the digital shift that we are witnessing in the information society. The following conceptual relationships are all affecting the topic of selection for long term archiving, irrespective of the heritage subdomains, and they should be taken on board in the guidelines.

a. physical/digital/hybrid
During the workshop the remark was made that although the focus of the selection guidelines is on born digital cultural heritage, we shouldn't neglect the relationship between the digital and the physical. A lot of digital objects that result from the process of digitisation of physical objects are unique and cannot be recreated at a later stage. Physicality is not a static condition. Examples mentioned during the workshop were digital images of paintings before their restoration, newspapers that are too brittle to digitise again and art installations, that are a digital/physical hybrid. It would be good if the guidelines didn't make a strict distinction between born digital and digitised cultural heritage.

b. product/process
In the world of collecting and curating physical objects, the focus of libraries and museums is quite strongly on end products (books, journals, objects), while at archives and archaeology the documentation of the process that leads to the end products is vitally important. In the workshop it was concluded that in the digital world, because of the dynamics in the curation of digital information, the data about the processes are getting equally important across all domains. The recording of activities and context that surround digital objects can be done in text (e.g. in metadata or in e-mails) as well as images/AV-recordings.

An example that was discussed more in depth at the workshop was research data. The researchers in the workshop mentioned that they would prefer that memory institutions give priority to archiving raw research data over final interpretations (e.g. reports or articles). From their perspective, the raw data and the metadata about the process should be available from the same instance. They want to be able to process that data in their own personalised environments. This creates new metadata and new interpretations. This led to the conclusion that selection and archiving should not be considered as an end point, but as a starting point for adding new meanings to archived resources. This was a confirmation of what is already mentioned before, that memory institutions need to take the full chain into considering when selection for archiving, from production to providing access and adding value after the archiving. However, it was agreed that the guidelines should stay focussed on selection, and not on providing or improving access.

In conclusion, the whole group agreed that preserving the context of digital cultural heritage is paramount. Digital resources that are being selected for archiving cannot be considered in isolation. Another example given was the archiving of digital scientific articles. The web references in these
articles (which serve as context information) need to be taken into account when archiving the article. Archiving should be treated from a holistic point of view.

c. institutions/communities/individuals

Archiving digital resources is not the exclusive responsibility of memory institutions. Private companies, communities and individuals are also playing an important role in the archiving of digital resources. There are several aspects here that affect selection policies by memory institutions:

• there are digital archives being created by individuals, either locally or on the web using personalised services of ICT-providers. The digital archives can be about their private life (e.g. photographs), or about topics or issues that they have an interest for. Many highlights in the physical collections of memory institutions are based on private collections, that were later donated to the library, archive or museum. The same is true for digital archiving. Memory institutions should have a keen eye on digital collections being managed by companies, communities or individuals, which may be incorporated in the collection of a memory institution at a later stage.

• As for the selection of digital content produced by communities, especially in the context of social media, memory institutions should make a clear distinction between three functions: a) the social medium as a platform (Facebook, Twitter) which results in big data to be archived; b) the community forum, which is constituted of narratives by professional or thematic groups; c) content produced within a community (e.g. fan fiction communities that write alternative endings to literature). It would be good if the guidelines state explicitly that these distinctions require different selection policies and preservation strategies. We should mention however, that there was no consensus among the workshop participants whether memory institutions should archive huge amounts of social media content (it was for example questioned whether it was useful that the Library of Congress archives Twitter).

• We also discussed the issue of personalisation. The participants in the workshop pointed out that private companies that act in the world of digital information (e.g. Google, Microsoft etc.) are getting less and less interested in the content that individuals handle. Their business models are built on personalisation. In the PERSIST position paper about selection it was also already mentioned that the social media platforms are not about the content, but about the digital footprints we leave behind to represent our digital behaviour, on which marketing strategies are based. This means that memory institutions should be aware of the fact that huge information sources are created about individuals that interact with digital content. This refers back to the remark made before that documenting the processes becomes equally important as archiving the content. It would be good if the guidelines explicitly refer to the existence of this public/private nature of large scale information sources.