Managing containers, content and context in digital preservation: towards a 2020 vision

Simon Tanner; King's College London; London; United Kingdom.

Abstract

20/20 hindsight is perfect, yet looking forward to 2020 the future is not yet so clear as to what it holds for digital preservation. Will we look back to now from 2020 and see a digital dark age or the beginning of a golden ambient intelligence environment? This paper will look at a classic information management metaphor, the information container and extend it to identify the challenges facing digital preservation as we move from managing information containers*, through* content *to information* context*. This paper will then discuss the research agenda and challenges that need soon to be faced and resolved to move our community forward, navigating through the Semantic Web towards the ambient intelligence context sensitive environment: where digital information becomes a ubiquitous process of perception and communication.*

Introduction

The digital birth and conversion of cultural content into bits and bytes has opened new vistas and extended the horizons in every direction; providing access and opportunities for new audiences, enlightenment, entertainment and education in ways unimaginable a mere 15 years earlier. Digital libraries have a major function to enhance our appreciation or engagement with culture and often lead the way in this new digital domain we find ourselves immersed within. Yet looking forward 15 years to 2020 the future is by no means clear as to what it holds for digital preservation. Will we look back to now from 2020 and see a digital dark age [1] or the beginning of a golden age of the ambient intelligence environment?

This paper uses a classic information management metaphor, the information container (as conceived by Suzanne Briet) and extends it to identify the challenges facing digital preservation as we move from managing information *containers*, through *content* to information *context*. This paper will then discuss the research agenda and challenges that need soon to be faced and resolved to move our community forward, navigating through the Semantic Web towards the ambient intelligence context sensitive environment: where digital information becomes a ubiquitous process of perception and communication.

We are now trying to move on from managing *containers* and *content* to managing *context* and it is proving to be an even larger and historically more difficult challenge to overcome. The Semantic Web views a future in which information is given well-defined meaning, better enabling computers and people to work in cooperation. The infrastructure of the Semantic Web would allow machines as well as humans to make deductions and organize information. We have moved from managing *containers*, to *content*, to *context* and at each stage the volumes of data and the complexity of the information domain has grown exponentially.

Managing Containers, Content and Context

Containers

Memory institutions have historically focused upon archiving, managing and preserving what can be termed *containers* of information: whether boxed letters, reports, documents, paintings, film or photographs. These physical, primary carriers of recorded information and knowledge content were the main focus of efforts to enable description and discovery by archivists and librarians in the past. In short, the route to acquiring the content of these documents was inherently wrapped up in a search for the container.

Suzanne Briet ("Madame Documentation") was an important French Documentalist. Her seminal work *Qu'est-ce que la documentation?* in 1951 [2] presented a uniquely strong attribution of containers as having a cultural origin and function and is expressed in examples of documents that include the, by now famous, antelope example. Briet was a follower of the founder of European Documentation, Paul Otlet, but she understood the cultural origins and functions of ideas such as *science* and *culture* more in the context of capitalist economies and the global development of the post-war period rather than Otlet's information utopia reminiscent of Victorian age great expositions. Briet was, to use a modern analogy, more cyberpunk than technocracy oriented. The important element of her work was the intrinsic importance of containers to the content and context they carried – something that has sometimes been mislaid in our digital domain.

Managing containers must not be perceived as to disparage what has been achieved over centuries of archival and information management. A container by its very existence assumes properties of content and context. Just as architecture is the management of space where walls, ceilings etc create boundaries in which the space may be utilized and understood; so to a container (bound volume, framed picture, photographic plate, printed map etc.) has an essential role in managing and using its content and understanding its context.

Managing containers as an achievable archival goal has led to a biblio-centric view where the containers are described and the *contents* and *context* of the information containers inferred from indexes, catalogues, classification schemes or collection management. The advantage of this is that at its most granular level this provides an extremely high level of discovery enabling a single letter in an enormous archive to be found easily. In reality, most archives are not this well described because it costs large amounts of money to get to this level of granularity for a very large collection. Frequently, the Archivist is the repository of this granularity of knowledge and the recorded discovery tools to containers are often described at the box level at best. Thus, discovery of the letter requires discovery of its box and then searching that box by hand to find the physical letter. In no way could this discovery be deemed as managing much content. Context is usually well represented in Archives. In this example, context will tend to be limited to the item in relation to its collection peers and that derived from discovery information.

In digital preservation terms we have been very successful in developing methods for preserving digital containers. Refreshing media and emulation to a certain extent address the issues around preserving the bit stream. However, migration techniques actively destroy and reincarnate the container in an attempt to preserve content and usability. Content is naturally given more prominence in digital preservation and as we will see next is clearly where most of our efforts focus in the digital domain.

Content

Otlet stated, ‘the book is only a means to an end. Other means exist and as gradually they become more effective than the book, they are substituted for it’ [3]. Otlet showed great foresight and realized that general users would become format agnostic and mainly focus upon the content not container. ‘To some extent, the “shapes” of the containers of information have been retained in a virtual world: e-journal pages look just like their print ones… But in a world where information and content increasingly are unbound from containers, the containers cannot act as guides… The second pattern to emerge from the twilight is the rapid and widespread reduction of content and institutions to much smaller units of use and interaction than in the past’ [4].

Managing containers is something Archives and other memory organizations have done very well indeed, but the growth in computing use from the 1960s onwards propelled archivists and other information workers into the direct management of content. Megill [5] points out that the information an organization needs to keep for re-use, that is worth sharing, managing and preserving to function effectively is the ‘corporate memory’. To keep, store and release this information in a timely fashion is the desire of most archival repositories.

Content held in digital form is in danger, not because the container is inherently fragile or flawed, but because there is a continually accelerating rate of replication, adaptation and redundancy of hardware, software and data formats and standards which may mean that the containers bit stream may not be readable, interpretable or usable long into the future. All data requires an element of decoding before it is recognizable and usable in a computing environment, even if open data standards are used. We take this automatic decoding for granted until we try to read a word processing file from 10 years ago and find that none of our current systems or software has any idea what the bit stream means without significant coaching or expert help.

Managing content on the Web throws up other issues for resource discovery and Web Archiving. The aptly named ‘Deep Web’, those massive resources missed by search engines due to being in a database or other non-harvested format, remains an often unnoticed problem to be resolved. Ironically, these may be very content rich and from a reliable publicly funded source such as an archive, library or museum. Michael K. Bergman and Bright Planet have issued a White Paper [6] that estimates:

- Public information on the deep Web is currently 400 to 550 times larger than the commonly defined World Wide Web.
- The deep Web contains 7,500 terabytes of information compared to nineteen terabytes of information in the surface Web.
- The deep Web contains nearly 550 billion individual documents compared to the one billion of the surface Web.

Effective techniques are needed for content and resource discovery. The Open Archives Initiative Protocol for Metadata Harvesting (OIA-PMH) is an efficient mode for metadata exchange and relies upon a minimum requirement for unqualified Dublin Core to enable effective cross-walking of resources. Of course, Dublin Core has such a basic structure that many complex resources are simplified so much that hierarchical richness in their encoding is lost at the point of cross-walking. Google, and other search engines, now uses OAI-PMH to harvest information and thus help repositories to bring their content to the surface.

Context

‘To see we must forget the name of the thing we are looking at’ Claude Monet.

Imagine a visit to Giverny in France. Whilst there, your portable device (which understands your location, preferences and interests) offers information to questions such as “tell me some historical information about this village and anyone famous who lived here”; “places to buy Monet souvenirs near where I am now”; “French gardens and painting”; “where did Monet live and what colour is his house” or “why did Monet paint flowers and gardens”. In a truly context sensitive world, widely divergent user needs would be supported when seeking for this sort of information. It is interesting that these questions are not unusually difficult in a human mediated environments (libraries, archives), but in a non-mediated digital environment they are not answerable unless very rigorous description and context has been provided. It is at the point of managing context that our digital plans have provided somewhat frugal results and frustrated our aspirations.

‘We do not see things as they are; we see things as we are’ The Talmud [7].

Finding a known object is always going to be easier than finding a range of previously unknown pertinent objects and if the starting perspective of the searcher is unknown because of diversity (age, education, language, etc) then making a resource findable when it might be text, audio, video, 3D, geographic, database or image based is a challenge to any digital repository. In a known case (e.g. Monet’s paintings of Giverny), searches can be constructed by inexperienced users that will almost certainly result in satisfactory retrieval. It is when the user knows only the field of enquiry, and not the precise resource, that search engines are very much less useful. Metadata and tools for resource discovery are needed to allow users to locate the items they seek, whether they know of their existence or not.

Because many of us are now of a generation who have grown up with computers as ubiquitous to our lives we forget that before their widespread use managing content was limited and very difficult; whilst managing context was resource hungry, time consuming and tended to reflect the narrow concerns of the organization archiving the content. Well, computer use for storage and manipulation has biased the equation towards managing content, especially in terms of volume and for textual resources. Otherwise, we still find ourselves managing containers and sometimes content but rarely is context any better managed than it was in the pre-computer archival record. Ted Nelson's 1965 aspiration for his Xanadu system, in which all the books in all the world would be 'deeply intertwingled' has still not been fully realized. In other words, we understand the principles and the need quite well, have applied it where possible, but rarely are the resources or infrastructure actually available in the digital domain to make contextual information widely shared, usable, robust and powerful.

Scholars have 'hailed the signal purpose of archives and special collections to preserve the context in which information arose or was fixed, used, or collected.' [8]

The Semantic Web and Ambient Intelligence Environment

With the dawn of digitisation came the opportunity, firstly for printed sources but latterly for other modes of information carrier, for *content* to not be inferred but directly managed, preserved and utilised. This move has provided a challenge for information workers and users alike with a concomitant information deluge where sorting out the useful from the chaff becomes ever more difficult. Studies by University of California at Berkeley [9] show that the United States produces about 40% of the world's new stored information, including 33% of the world's new printed information, 30% of the world's new film titles, 40% of the world's information stored on optical media, and about 50% of the information stored on magnetic media. This explosion in information, services and resources, whether appropriate to the users needs or not, all consume attention. Information has to be selected or discarded, read or not read, but it cannot readily be ignored. The actual downside of the information explosion is a deficit of attention, known more popularly as 'information overload'.

We have moved from managing *containers*, to *content*, to *context* and at each stage the volumes of data and the complexity of the information domain has grown exponentially. The architectural components include semantics (meaning of the elements), structure (organization of the elements), and syntax (communication). The use of RDF (Resource Description Framework) and XML (eXtensible Markup Language) are essential elements of this approach. The International Council of Museum's common extensible semantic framework, CIDOC Conceptual Reference Model (CRM) provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. It is an excellent example of the power of semantic approaches to express common concepts in basic categories that will have real longevity in use.

The ambient intelligence environment goes even further and the term relates to a new way of conceiving the role of information technology, a role in which the digital environment is aware of a person's presence and context, and is responsive to their needs, preferences, and desires. As computing becomes ever more ubiquitous and connectivity of almost every mobile device with advanced, robust ad-hoc networking technologies possible then the infrastructure for the ambient intelligence environment may be established. What are missing from this equation are the adaptive user-system interactions and so-called social user interfaces, to enable the digital environment to act on our behalf to enable entertainment, education and enlightenment. 'These context aware systems combine ubiquitous information, communication, and entertainment with enhanced personalization, natural interaction and intelligence' [10].

Digital preservation, as a discipline, looks to follow a similar metaphoric path. We are currently managing digital containers quite well. The files and data sets that make up our digital resources are managed through reformatting and migration methodologies that, whilst not perfect, are pretty well understood. However, managing and retaining content and context is far more challenging. Preserving the content and experiential element of digital resources is currently in the domain of migration and emulation and there is still much work to be done. This author would further suggest that context digital preservation is not achieved in any meaningful way at present and resources are bifurcated and separated by the very infrastructure that seeks to create modes of connection and contextualization.

Barriers

There are a number of major barriers in front of the community inherent in this continuing transition from managing *containers*, to *content* and *context.*

Attempts at full interoperability between varied archives, systems and standards, and between communities have not yet succeeded, and seem unlikely to succeed in the near term. Even assuming technical hurdles can be overcome there are political issues of: control; resources; legal frameworks; regional, national and international community differences to be overcome. The goal of the world-wide semantic web is probably unreachable while the issue of interoperability still remains as the biggest sticking point of all.

The major unresolved issues in the transition revolve around money, infrastructure, scalability and sustainability. Frankly, managing content and context in digital repositories is a large and unfunded mandate that has been forced upon the community because of perceived user demand and the short timeframe in which action must be taken or the resource will not just be unmanaged but lost to the future. IT in archives is no longer showing the immediate return on investment delivered in the 1980s and 1990s, such that future developments will not necessarily instantly save staff time or reduce costs. The current benefits from technology for archives are improving resources, processes and services, not replacing the human factor. Of course, the issues of sustainability and scale become paramount once significant investment has been made and there is an undercurrent of dissatisfaction regarding the sustainability and scalability of digital technology regarding issues of preservation and continuing access to resources.

Perspectives of value and the incentives to contribute are skewed in relation to digital preservation. Many information users would like the Semantic Web where information is intuitively available. Yet few show any interest in expending effort to document information for use by others – time expended recording contextual information and improving accessibility is time not spent on new activity. For scholars, in particular, the rewards are for publication not information management or archiving. The ideals of open scholarship date back to at least St. Augustine, yet issues of control and ownership still stifle sharing content and thus hinder the contextual recording that would fully enable the Semantic Web. The user assumption is that the archive will take care of all this, but please don't ask for any resources or funding to do this!

Debate about money and infrastructure is predicated upon an understanding of the community who will benefit from such investments. Such conversations are thus hindered by a lack of metrics or evidence base to show a clear understanding of their 'designated communities' needs and desires. The business models being implemented are not exciting the consumer, whether academic or from the general public. Some of the quasi-commercial outsourced data repositories order books look very thin compared to investment and the financial failure of some of them is a real prospect. Further, we have yet to engage in a fully fledged quantitative risk management and rely upon assertions, case studies and qualitative consequences of digital loss to justify our activities. This leaves those seeking to justify activity, especially for extending metadata or contextualization for resources, with a weaker argument as it is not often backed by a strong evidence base. As stated by a respondent to the Digital Preservation Coalition's 2005 Survey: 'costs (and indeed all kinds of resources) are very difficult to quantify and forecast reliably. There remains a lack of standards and benchmarks, and this makes it hard to compare ourselves (and costs) with other organizations' [11].

Research Agenda

The current research activity in digital preservation regarding the management of containers, content and context generally seeks to address he inadequacies of current strategies and the means to deal with increasingly complex and bifurcated digital entities. Three reports, 'It's About Time' [12], 'Invest to Save' [13], and 'Mind the Gap' [11] propose agendas for digital archiving. A meeting in November 2005 in the UK on Digital Curation also sought to set the research agenda for the next decade [14].

The research challenges are broadly defined across 5 general groupings:

1. **archival repositories** - technical architectures, models, format repositories;
2. **archival collections attributes** – metadata, interoperability, context-aware digital entities, function and behaviour documentation, automated metadata creation;
3. **archiving tools and technologies** – salvage and rescue, media, formats, storage, accelerated aging, anomaly detection, multilingual entities, and automation;
4. **strategy, policy, economic, and risk management issues** – intellectual capital, authenticity and information quality, scalability, repurposing,
5. **metrics, evaluation, performance, and effectiveness** – modelling preservation processes, collection completeness, acceptable loss, quantifiable risk, cost benefit analysis tools,

Hedstrom identifies 'that human labor is the most costly element in digital preservation and one that is likely to increase, while storage and processing costs continue to decline. Therefore, there is a premium on developing methods that reduce the amount of human intervention in digital archiving processes' [15]. This is clearly important, but the bigger goal may not be to remove human intervention but to gain greater value from it. One way this could be achieved would be to focus on ways humans can augment contextual management rather than spending time managing content and containers.

More research attention should also be focused on metrics and quantifiable factors that deliver cost models against benefits, risks and values of digital objects. 'Survey after survey conducted over the past five years provides a bleak picture of institutional readiness and responsiveness. Why this lag in institutional take-up? In part the answer lies in the fact that most of the attention given to digital preservation has focused on technology. This emphasis has led to a reductionist view wherein technology is equated with solution, which in turn is deferred until some time in the future when the technology has matured' [16].

Metadata is clearly of vital importance to the Semantic Web and must be more widely deployed and very much more scalable than it is now. At the very least, a clear definition of its value must be promulgated across the community and creator, mediator and user need to understand their roles and respect the benefits concomitant in metadata creation. Metadata will certainly be captured closer to the point of resource creation. Further research to discover means and modes to bring the 'designated community' and the service providers closer together are essential.

At the discovery level, metadata must be developed that allows descriptions of content and context that will be understood and processable by machine to machine interactions. Deep semantics and domain ontologies plus taxonomies need to be fully populated and linked. This will enable tools for greater levels of automated metadata creation, capture and update to be widely utilised. The goal of enabling more metadata to be inferred automatically from the resource characteristics will ensure that where human intervention is required it will deliver greater value.

2020 Vision

'Trying to predict the future is a mug's game. But increasingly it's a game we all have to play because the world is changing so fast and we need to have some sort of idea of what the future's actually going to be like because we are going to have to live there, probably next week' [17].

What can possibly be said about the state of digital preservation and the working theme of this paper of managing containers, content and context in the year 2020 that will not be woefully and embarrassingly incorrect? Global disaster notwithstanding, it would not be too great an assumption that devices will become ever more connected and powerful such that the environment in high connectivity countries will become "intelligent". This assumes that the device will be able to identify: geographic location, user identity and authentication, personalizations and preferences; and use these to interact with other devices. If this becomes achieved, as seems likely, then it would be a tragic waste of connectivity if that device was not also able to interrogate local devices and the wider network to find information and resources of interest to its pre-defined (and possibly self-taught) set of owners preferences, desires and needs.

The day may come when a tourist wonders into a cathedral and has the local tour automatically available to them in their own language, plus: images and information on the stained glass to high to view, video of famous ceremonies carried out, the historic plans with a 3D visualization of what the cathedral may have looked like 200 years previously, full text of historic and literary references to the cathedral, a list of people buried, baptized or married there, choral works performed; and the list could go on.

The above scenario would be quite an astonishing feat when we consider the state, not of the current technology, but of the strategic and research agendas for digital repositories and archiving. Without massive resource interoperability, context sensitive digital objects, deep metadata and ontologies with concomitant supporting business models then this scenario will remain just a dream.

We must not assume that technology will solve our information problems – in that direction lies a frustrating wait. It is time to stop avoiding the clearest conclusion possible from the current state of the digital domain: higher levels of human intervention are needed and are in fact desirable to enrich and make more valuable all our digitization and preservation efforts to date and for the future. Without this commitment of time and energy the chances of a vibrant, open information environment are slim. In such a vacuum, commercial interests will bifurcate the market into its smallest common denominator with all its context stripped away - much in the way that medieval manuscripts are fiscally more valuable divided up for sale into individual leaves (or even single letters). If we lose sight of the value of investing in context, and the incredible possibilities that managing context presents, then we may very well look back from 2020 and regret that we did not grasp this opportunity.

References

[1] M. Deegan and S. Tanner, "The digital dark ages: digital preservation" Library and Information Update, May 2002.

[2] S. Briet, "What is Documentation?" Translated by Ronald E. Day and Laurent Martinet. [Qu'est-ce que la documentation? Paris: Éditions Documentaires Industrielles et Techniques (EDIT), 1951]. at /www.lisp.wayne.edu/~ai2398/briet.htm, last accessed 03/10/2006

[3] Otlet, 1934, quoted in Rayward, 1994, 244).

[4] C. De Rosa (Contributor), L. Dempsey (Contributor), R. Limes (Contributor), L. Shepard (Contributor), A. Wilson (Editor), "The 2003 OCLC Environmental Scan: Pattern Recognition: A Report to the OCLC Membership", OCLC, 2003.

[5] K. A. Megill, The corporate memory: information management in the electronic age, Bowker Saur, 1997.

[6] M. K. Bergman, "The Deep Web: Surfacing Hidden Value", at www.brightplanet.com/technology/deepweb.asp, last accessed 03/10/2006

[7] "The Talmud", The Jewish Virtual Library at www.jewishvirtuallibrary.org/jsource/Talmud/talmudtoc.html, last accessed 03/10/2006.

[8] A. Smith, "In Support of Long-Term Access" chapter in "Access in the Future Tense", Council on Library and Information Resources, April 2004.

[9] P. Lyman and H.R. Varian, "How Much Information 2003?" University of California at Berkeley, at www.sims.berkeley.edu:8000/research/projects/how-much-info-2003/, last accessed 03/10/2006

[10] ITEA Ambience Project at www.hitech-projects.com/euprojects/ambience/, last accessed 03/10/2006

[11] M. Waller and R. Sharpe, "Mind the Gap: Assessing digital preservation needs in the UK", Digital Preservation Coalition, 2006.

[12] It's About Time: Research Challenges in Digital Archiving and Long-Term Preservation, Final Report, Workshop on Research Challenges in Digital Archiving and Long-Term Preservation, April 12-13, 2002, sponsored by the National Science Foundation, Digital Government Program and Digital Libraries Program, Directorate for Computing and Information Sciences and Engineering, and the Library of Congress, National Digital Information Infrastructure and Preservation Program, August 2003.

[13] Invest to Save: Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation, prepared for the National Science Foundation's (NSF) Digital Library Initiative and the European Union under the Fifth Framework Programme by the Network of Excellence for Digital Libraries (DELOS), 2003

[14] Digital Curation and Preservation: Defining the research agenda for the next decade, 7-8 November 2005, Warwick. Drivers and Barriers Session Report.

[15] M. Hedstrom, "Research Agendas Set Course for Digital Archiving and Long-Term Preservation" RLG DigiNews, December 15, 2003, Volume 7, Number 6, at www.rlg.org/legacy/preserv/diginews/v7_n6_feature2.html, last accessed 03/10/2006.

[16] A. R. Kenney, "Collections, Preservation, and the Changing Resource Base", chapter in "Access in the Future Tense", Council on Library and Information Resources, April 2004.

[17] D. Adams, "Predicting the future", H2G2 Guide ID: A216433, at www.bbc.co.uk/dna/h2g2/A216433, last accessed 03/10/2006

Author Biography

Simon Tanner is Director of King's Digital Consultancy Services (KDCS) at King's College London. KDCS provides research and consulting services specializing in the information and digital domain for the cultural, heritage and information sectors.

Tanner has a Library and Information Science degree and is an independent member of the UK Legal Deposit Advisory Panel and Chair of its Web Archiving sub-committee. Tanner authored the book, Digital Futures: Strategies for the Information Age, with Dr Marilyn Deegan.