The Preservation of Complex Objects

Volume 1

Visualisations and Simulations

2012

Janet Delve, David Anderson, Milena Dobreva,
Drew Baker, Clive Billenness, Leo Konstantelos

POCOS

Preservation of Complex Objects Symposia

The Preservation of Complex Objects

Series editors:
David Anderson, Janet Delve, Milena Dobreva

Volume 1. Visualisations and Simulations

Volume 1 editors:
Janet Delve, David Anderson, Milena Dobreva,
Drew Baker, Clive Billenness, Leo Konstantelos

Published by The University of Portsmouth.

ISBN 978-1-86137-6305

First published 2012.

Preface

Richard Beacham

Founder and former director, King’s Visualisation Lab, King’s College, London, UK

“And on the pedestal these words appear:
‘My name is Ozymandias, king of kings:
Look on my works, ye Mighty, and despair!’
Nothing beside remains. Round the decay
Of that colossal wreck, boundless and bare
The lone and level sands stretch far away.[1]

Over the past two decades there have been extraordinary developments, enormous interest, and stunning achievements in the use of digital technologies for humanities research. In particular, scholarship has become increasingly aware of the powerful new media for research, learning, and communication afforded by the creation of visualisations and simulations derived from complex digital 3D models of architecture and artefacts. The contents of this volume detail both the nature of the new research-based “artefacts” arising from such work, and of the tools and strategies required for their creation, ongoing access, and preservation.

The research team that I founded and directed over some 15 years, and which now as the “King’s Visualisation Lab” is based at King’s College London, has played a major role in these endeavours. Examples of the nature of its work and a few of the very many projects led by it feature here in accounts of the “Body and Mask in Ancient Theatre Project”, the “Roman Villa of Oplontis Project” (for which KVL created the detailed 3D model of the villa), the creation of the online “virtual world” THEATRON, and the “London Charter” initiative.

As major “stakeholders” in this ever-expanding, deeply challenging, and endlessly intriguing new field of scholarly exploration and development, we have – conscientiously and inevitably – been subject to what I will term the “Ozymandias syndrome”. This is an acute awareness of the vulnerability of our own meticulously researched and lovingly crafted creations to being, like Shakespeare’s “insubstantial pageant”, “melted into air, into thin air”. “Virtual” artefacts, whatever the source of their inspiration and however grand or impressive their realisation, are, as detailed in the accounts which follow, like Ozymandias’ proud “works” subject to nemesis: their assertions of grandeur and enduring significance mocked by their all too frequent disappearance into boundless and bare cyberspace.

It is of course deeply ironic that these labours of love and scholarship, intended to salvage endangered or lost cultural heritage from the ravages of time and mutability should be themselves in such imminent peril. But they are. And indeed, for KVL there is an added poignancy, since from the first we focussed upon the virtual replication and “salvation” of one the most ephemeral examples of cultural heritage; theatrical settings and stage production. 3D modelling of the fleeting rudiments of theatre history is an ideal and unprecedented research medium for evoking such core and definitive expressive elements of the art of the theatre as time, space, movement, scenic elements, visibility, and sequence, in a manner that traditional scholarship could never achieve. It makes these qualities visible, “tangible” and subject to analysis and comparative study in a way previously denied to scholars.

[1] “Ozymandias”, lines 9-14, P. B. Shelley. First published in *Rosalind and Helen: A Modern Eclogue; with Other Poems* (London: C. and J. Ollier, 1819).

But now we face a crisis in the preservation of such new creations. Humanities scholars have been encouraged and supported to use technology both for research and for its publication/dissemination; but there is a yawning "gap" – a discontinuity – now between the means for creating such work (which we are increasingly good at!) and the means to preserve and ensure its long-term sustainability; an area in which we are failing lamentably. The contents of this book identify and document the extent and urgency of the problems, and attempt to suggest, tentatively, how they might be addressed.

Over a great many centuries we have learned how to preserve printed data in books or manuscripts for posterity, and of course we strive – albeit inadequately – to preserve what remains (if anything) of the physical artefacts which are now becoming the objects of our virtual replications. Why do we not have the same expectations and aspirations for digital creations? After all, in, for example, the case of virtual 3D models of vanished architecture and objects, increasingly these are not merely the illustrations of text based/ text delivered research; they *are* the research itself, and in the case of "virtual worlds", also constitute the very venue and occasion in which such research can be enabled to take place. And yet, as detailed in the account by Drew Baker and other contributors to this volume, resources, whether financial or technical, are not normally provided to preserve such research. External funders (who have accepted the promise and the premise of new technologies, and generously funded our endeavours, are understandably *not* willing to provide recurrent funding to sustain it, which they view as the responsibility of institutional infrastructure. But this is an understandably difficult message for institutions to take on board, especially at this time of austerity.

Thus the challenges we must address are complex, and the battles to be fought require both vigilance and engagement on several fronts. On the one hand we face the slings and arrows of the extensive technical and methodological obstacles and shortcomings identified and discussed at length here. On the other, we must confront the constraints faced by our institutions. As contributors to this volume have noted, in addition to software and hardware issues, it is also essential always to have due and vigilant regard for the human factor: the so-called "wetware". Ultimately, and obviously, both the scholarly input and research output as these are modulated through the media of hard and soft ware, are the product of human thought, judgement, choices and interpretation – what Drew Baker coined the term "paradata" to designate. It is important to note and to stress that much of the pioneering and most influential work in the application of virtual technologies to research has been led initially by "traditional" scholars, and certainly this was true in the example of the theatre historical work of KVL.

The work which such scholars had undertaken deploying the "orthodox" tools of their discipline prior to their engagement with digital media lent confidence and credibility to the quality of their new and less conventional scholarship, and thereby served *pour encourager les autres*, to follow in their path. Moreover, such academic pathfinders were themselves conditioned readily to discern just how valuable such new technology and the methodologies enabled by them could be in addressing the sort of fundamental issues and questions with which their discipline – for example that of theatre historical research - had from the beginning been profoundly engaged. One of the still evident challenges hindering fuller take up of digital humanities in general is to bridge the gap between experts in digital technology and experts in particular academic disciplines, and by minimising if not entirely effacing such lines of demarcation allow a richer scholarly synthesis to arise, and to be nurtured and promulgated.

If I may be permitted a personal note, hosting the POCOS symposium of which this volume is an outcome was one of the last activities I was privileged to undertake as the Director of KVL. My own role and functionality, figuring from the beginning in the work of KVL as its most central and durable piece of rather complex "wetware", thus regrettably came to an end. As a team which, throughout its existence, has with considerable success and gratifying recognition travelled and trafficked in "envious and calumniating time"[1] while trying to preserve (if only virtually) both the endangered or vanished objects of its research, and the digital surrogates arising from its members' scholarship, KVL looks forward to continuing its guardian role to ensure future simulations and visualisations do not disappear into "the lone and level sands".

[1] *Troilus and Cressida*, Act 3, Scene 3.

Acknowledgements

Preservation Of Complex Objects Symposia (POCOS) has been funded by the JISC Information Environment Programme 2009-11.

Contents

Current Digital Preservation Approaches and Practices in Simulation and Visualisation Projects

Case Studies and Discussion Topics

Introduction to POCOS e-Book 1: Preserving Visualisations and Simulations

Janet Delve

Future Proof Computing Group, School of Creative Technologies, Eldon Building, University of Portsmouth, Portsmouth, PO1 2DJ, UK

Background to POCOS

The preservation of complex materials and associated environments presents the Digital Preservation (DP) community in general and the JISC community in particular with considerable intellectual and logistical challenges. While many of the techniques that have been developed within the context of migration-based DP approaches are of continuing value, others cannot be applied so well given the extra complexity presented by, for example, interactive videogames. Recent work undertaken in the Planets and KEEP projects has shown that the problems involved in preserving such materials and their associated environments, while substantial, are by no means intractable, but in order to continue to make progress in this area it is important to engage and energize the wider DP community. A vital aspect of this process comprises articulating the state of the art in 1. *Simulations and Visualisations*; 2. *Software Based Art* and 3. *Gaming Environments and Virtual Worlds*. This encompasses exploring with international experts the research results achieved so far across each of these domains; presenting coherent pathfinder solutions; and clearly signposting areas where work remains to be done. A further step is to synthesize key findings across all three areas and emphasize synergies that can be built upon, and to disseminate these to the various stakeholder communities.

These are the principal objectives that POCOS addresses and POCOS partners are well-placed to tackle the problem space, with the University of Portsmouth as overall coordinator bringing research and technical input from KEEP; the British Library supplying project management and research expertise from Planets, King's Visualisation Lab bringing their specialist visualisation and simulation knowledge and experience; The Humanities Advanced Technology & Information Institute giving their specialist Software Based Art expertise from Planets; and Joguin sas supplying graphical input, and technical experience from KEEP.

So, in a series of three symposia presented across the UK:

- *Simulations and Visualisations* organized by the King's Visalisation Lab (KVL) at King's College London on June 16th and 17th 2011;
- *Software Based Art* organized by The Humanities Advanced Technology & Information Institute (HATII), the University of Glasgow, at the Lighthouse, Glasgow on October 11th and 12th 2011; and
- *Gaming Environments and Virtual Worlds* organized by the Future Proof Computing Group, the University of Portsmouth at the Novotel Hotel, Cardiff on January 26th and 27th 2012.

POCOS brings together the leading researchers and practitioners in each field to present their findings, identify key unsolved problems, and map out the future research agenda for the preservation of complex digital materials and their related environments. The fundamental task to be faced during these symposia lies in presenting specialist material of great technological and organizational complexity in a lucid, cogent, relevant and approachable manner so as to engage UK HEI researchers and practitioners in a wide variety of disciplines, as well as reaching those further afield in, for example, commerce, industry, cinema, government, games and films classification boards, and healthcare. There is also the added concern that the specialists in each field may not necessarily be aware of general trends in DP, and vice versa. Similarly any differences in terminology might need carefully addressing. Hence, clarity of expression and good communication is thus paramount throughout all the exchanges and discussions.

To this end, there is a series of three e-books, one for each symposium output plus any additional salient material, available from the POCOS website[1]. There is also a final compendium book covering all three symposia, together with a set of pathfinder solutions. This e-book is the first of the three, and as such starts off by considering some fundamental issues, such as the nature of complex digital objects.

What is a Complex (Digital) Object?

An essential first step when considering the nature of complex digital objects is to recognize that there are multiple layers of difficulty encountered when attempting to analyse them. These layers could be superficially likened to Georg Cantor's "levels of infinity"[2] in terms of mapping out the size of the problem space to be analysed. The first "level of infinity" is that of detail: the problem of drilling down through many layers of technical elements, showing levels of interconnectedness both within digital objects themselves, and also with their technical environments. An example of such a challenge is that of preserving video games under emulation that was the subject of a broad, systematic, in-depth study in the EC KEEP project[3] (Pinchbeck et al., 2009).

Similar advances have been made in the EC Planets project[4] where running interactive digital art under emulation and virtualization was examined in depth and scientific experiments conducted within the Planets Testbed environment. Analysing and mapping such a great level of detail is not just confined to emulation. The migration community has responded to the task of recording each aspect of a complex digital object by developing ontologies of significant properties, and the Planets project played an important role in both conducting and disseminating this research (Dappert & Farquhar, 2009). However, significant properties under migration encompasses not only the "level of infinity" concerning detail, but also another one to do with scale. Emulation also addresses the issue of scale as in practice it necessitates mapping out the necessary hardware, software, middleware etc. that makes up the technical environment of each complex digital object. The characterisation work in Planets, and technical environment modelling activity in KEEP thus represent important aspects of the state of the art in this problem space, and have provided a firm foundation from which to develop the area. So, from this springboard, how do we start to tackle the task of analysing the complex digital object *per se*?

The notion of the digital object is a mainstay of everyday life in mainstream digital preservation: indeed it is a concept that is fundamental to the way we approach this whole

[1] http://www.pocos.org/index.php/publications

[2] Developed at the end of the nineteenth century.

[3] http://www.keep-project.eu/ezpub2/index.php Keeping Emulation Environments Portable (KEEP) is a medium-scale research project which started on 1 February 2009 and is co-financed by the EC's 7th Framework Programme (ICT-3-4.3 Digital libraries and technology-enhanced learning priority).

[4] http://www.planets-project.eu/

domain using OAIS (CCSDS, 2009), PREMIS (OCLC/RLG, 2008) etc. Now, we can categorise an object as being atomic or complex: for example Hunter and Choudhury refer to "atomic or composite mixed-media digital objects" (p4). Another antipodal reference to complex digital objects comes from Somaya Langley[1] at the National Library of Australia's Gateway, who visited California in 2006 to study aspects of this subject area in three institutions (and incidentally came across the Media Art Notation System MANS that will feature heavily in the second e-book on Software Based Art). But it is really possible to separate digital objects into atomic and complex?

Let us say that there is an implication that an atomic digital object is a single file, and that this is synonymous with the notion of simplicity. But is that really the case? A single PDF file is often put forward as an exemplar of such a straightforward file, but the recent PDF 2.0 version can contain embedded 3D objects, so can it really be considered as atomic and 'simple'? So it might be a somewhat daunting task to rigidly categorize digital material past, present and future as either atomic or complex? During the symposia, the POCOS strategy was not to seek to impose definitions or standards on the proceedings, but rather to see whether any consensus emerged during the talks and breakout sessions. So given that general standpoint, how are complex objects regarded in terms of *Simulations and Visualisations*?

The Nature of Simulations and Visualisations

First it is important to note that *Simulations and Visualisations* are in a somewhat different category from *Software Based Art*; and *Gaming Environments and Virtual Worlds*; as the last two are each cognate disciplines in their own right: *Software Based Art* has dedicated artists, museums, techniques and commissioning procedures; and *Gaming Environments and Virtual Worlds* have their own games developers, games museums, conferences for the gaming community etc. *Simulations and Visualisations*, on the other hand, are amorphous techniques / outputs that are used in many different fields. The predominant domain in which they are considered in POCOS is archaeology, with the Kings Visualisation Laboratory, Kings College London leading the investigations. However, there are many other fields where such techniques and outputs are vital, and one suggestion that arose at the end of the first symposium was the use of simulations and visualisations in film, which led to the commissioning of a chapter on preserving film animations that will appear in the final book. There is one important topic, however, that arises from considering the items from an archaeological perspective: that of the *hybrid digital object*.

The definitions of complex digital objects given above were in the context of purely born-digital material. What arises in the mainly archaeological context that features in POCOS is the fact that many of the items are *composite*: *being part physical and part digital*, thus giving the DP community a *hybrid digital object*[2] to take into account. For example, there may be a part of a fresco from the Villa at Oplontis, together with a 3D rendering of some of its lacunae. This is of particular relevance when considering the different conventions already in place for preserving physical artefacts as opposed to their digital counterparts. And coincidentally, a similar situation occurs in *Software Based Art*

[1] http://www.nla.gov.au/pub/gateways/issues/84/story05.html

[2] See Susan Thomas, 2007, "Paradigm: A practical approach to personal digital archiving", The Bodleian Library, *http://www.paradigm.ac.uk/projectdocs/jiscreports/ParadigmFinalReportv1.pdf* for a discussion of hybrid digital objects in the context of personal archiving.

with physical components of all shapes and sizes (including live snails) linked to their digital counterparts.

The Book Contents

Alongside this introduction is also one by Drew Baker of the KVL who sets the scene for digital preservation from the visualisation / simulation perspective.

The first section then covers Current Digital Preservation Approaches and Practices in Simulation and Visualisation Projects, the first two chapters of which do not come straight from the symposium, but which are vital to this discussion: they concern the issue of preserving software. All complex digital material is by its very nature also software and much good work has already been done in this area. These two chapters by Neil Chue Hong and Brian Matthews et al respectively come from presentations at a JISC seminar in February 2011 on preserving software, and they provide both an excellent introduction to digital preservation in general as well as solid guidelines for all stages of preserving software, including case studies. We make no apology for the length of these chapters: the subject is complicated and involved and necessitates such a detailed and thorough exegesis.

We then move on to the specific problem space of archaeology with John R. Clarke's chapter on the Villa at Oplontis, which enumerates the issues surrounding the preservation of hybrid digital objects, and this theme is continued by Jenny Mitcham who delineates the practices and standards adopted at the Archaeology Data Service, together with case studies and exemplars. This section is rounded off by a chapter by Kenton McHenry et al in which a plethora of cutting-edge tools and techniques from the National Centre for Supercomputing Applications, University of Illinois at Urbana-Champaign are discussed in terms of their suitability for preserving 3D visualisations and simulations.

The next section: Case Studies and Discussion Topics, starts with a chapter by Hugh Denard on the London Charter: a fairly recent approach to preserving visualisations spearheaded by KVL and others that has now become standard. In his chapter Daniël Pletinckx explores the DP difficulties faced by cultural heritage institutions. Martin Blazeby presents a case study about the particular difficulties in preserving visualisations of masks used in Greek plays. The chapters on the discussion topics start with the session by William Kilbride of the Digital Preservation Coalition (DPC) on DP for preserving visualisations and simulations. Main themes that emerged here were the acknowledgement that emulation has now come of age as a suitable DP strategy to tackle preserving complex digital objects. In his chapter David Anderson summarises the findings from the layman's guide to the KEEP legal studies on how the law affects DP issues surrounding media transfer and emulation for multimedia works. The topic of information security led by Andrew Ball forms the basis for the next chapter. Lastly, Anouar Boulal et al present a formal technical approach which structures, disseminates and reuses complex digital objects of potentially any format. The applicability to existing scholarly publications systems is paramount for the eco4r project[1] in which these results were achieved. Finally a glossary appears at the end of the book to guide the unwary DP traveller through the minefield of acronyms…

We hope these research findings will help to stimulate debate on these topics, and look forward to continuing along this rich vein of research.

[1] http://www.eco4r.org. The eco4r project is a German Research Foundation (DFG) funded collaboration between Bielefeld University Library and North Rhine-Westphalian Library Service Centre (hbz).

Links to the Next Books

The second e-book is on Preserving Software Based Art and involves emerging issues such as the value of commissioned digital artwork, as well as others that have appeared in the first book, such as hybrid digital objects. The third e-book is on Preserving Gaming Environments and Virtual Worlds. Details will be available on the POCOS project website.[1]

References

CCSDS. (2009). Reference Model for an Open Archival Information System (OAIS) Retrieved from http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf

Dappert, A., & Farquhar, A. (2009). Significance is in the Eye of the Stakeholder. In M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou & G. Tsakonas (Eds.) Research and Advanced Technology for Digital Libraries. (Vol. 5714, pp. 297-308): Springer Berlin / Heidelberg.

Hunter, J., & Choudhury, S. (2006). PANIC: an integrated approach to the preservation of composite digital objects using Semantic Web services. *International Journal on Digital Libraries* 6(2), 174-183.

OCLC/RLG. (2008). PREMIS Data Dictionary for Preservation Metadata. Retrieved from http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

Pinchbeck, D., Anderson, D., Delve, J., Ciuffreda, A., Alemu, G., & Lange, A. (2009). Emulation as a strategy for the preservation of games: the KEEP project: Breaking New Ground: Innovation in Games, Play, Practice and Theory. Paper presented at the DiGRA2009, Brunel University. Retrieved from http://eprints.port.ac.uk/2714/

[1] http://www.pocos.org/index.php/publications

Laying a Trail of Breadcrumbs – Preparing the Path for Preservation

Drew Baker[1], David Anderson[2]

[1] King's Visualisation Lab, Department of Digital Humanities, King's College London, 26-29 Drury Lane, London WC2B 5RL, UK

[2] Future Proof Computing Group, School of Creative Technologies, Eldon Building, University of Portsmouth, UK, PO1 2DJ

Recent decades have seen an exponential increase in the use of computer systems in academic disciplines traditionally thought of as being non-technical. This has partly been driven by the opportunities which computational tools and techniques make possible, and partly by a perceived need to position research at the 'cutting edge' of technology. The use of computing is seen as enhancing research funding applications, and improving institutional profiles. The rapid uptake of computing has not, however, been accompanied by a commensurate attention to ensuring that digital-based research and deliverables are preserved in the long term.

The process of ensuring long-term access to non-digital material is well understood and while not always fully implemented, it is nevertheless relatively uncomplicated. The situation is markedly different when dealing with material that is wholly or partly digital. The preservation of, and provision for long-term access to, even the simplest digital object typically involves marshalling a surprisingly large number of technologies and specialized curatorial skills and tools, and no small degree of expense. In the case of complex digital objects, or hybrid objects, there are only a handful of centres in the UK or internationally who possess the knowledge or resources to tackle the preservation issues involved.

In this introductory chapter we will set out very briefly, and in fairly general terms, the overall preservation landscape as it applies to digital objects. We will try to give some sense of the main approaches available, and their strengths and weaknesses

Before looking at the principal preservation approaches currently in use, it is probably worth mentioning a fundamental difference between traditional objects and their digital counterparts. A standard textbook is capable of being read and understood by anyone who has sufficient intellectual capability, language skills, comprehension and so on, together with physical access to the material. The very same material in digital form requires something more – a facilitating layer of technology through which access to the content must always be mediated. It is usually possible to access the contents of digital files using more than one particular technology, one might use a Kindle or a PC for example, but direct unmediated access is not possible. The increasing digitization of material opens up very many opportunities for knowledge to be disseminated and accessed in new ways, and for new audiences, but the price of these potential benefits, is that without available technology there can be no access to information stored in digital form.

This opens up the question of what happens when the technology originally used to access preserved digital objects is superseded by new and incompatible devices. If some preservation action is not taken, the answer is simple, the old digital material will gradually be accessible on fewer and fewer platforms, and, when the last compatible device ceases to work, the digital objects will be inaccessible.

Two main techniques have emerged for retaining long-term access to preserved digital objects. The first of these, and by far the more commonly used, is migration, or format-shifting. The focus of this approach is the actual digital object to be preserved, and migration requires that preserved files are converted so as to be compatible with whatever hardware platform is available to access them. Thus files originally written in 1960 on a DEC PDP-1 machine and solely intended for use on that system may be converted to run on a VAX machine in the 1970s or on an iPAD2 in 2012. For simple file types the process of developing file conversion tools is not a particularly complicated one. However, there are a relatively large number of known file types in existence (somewhere in excess of 6000) and some of these are so intimately bound up with the particularities of their originally expected hardware that the conversion process becomes more tricky. So, in addition to potentially needing to develop thousands of individual file conversion programs each time a new computer system is invented, each of these must be checked very carefully to see the effect it has on the content of preserved files.

Even when a great deal of care has been exercised, it is not uncommon for files originally written for one computer system to appear slightly differently on a new system: colours, fonts, or pagination may be all altered, sometimes so slightly that detection is hard to guarantee. While in some cases, this may not be important, in others it may crucially alter the intellectual content or meaning in ways which are not systematically predictable. Moreover, these conversion errors, gradually accumulate as a file moves from one system to the next leaving open the very distinct possibility that over a relatively short time it will be nearly impossible to view digital material in its original form. Furthermore, the substantial expense involved in digital storage is such that within an institutional context, it is by no means assured that a copy of the original files will be preserved after conversion has taken place. In very many cases this will mean that post-conversion there will be no route back to the original material.

Migration is a technique best applied to simple file types, such as ASCII, or PDF 1.0. When faced with more complex digital objects, such as modern video games, virtualizations, or computer-based artworks, migration is not a viable option particularly within an institutional context. Large scale preservation activity is always expensive, but the time, equipment and expertise required to migrate (or port), for example, interactive video games from one hardware system to another and to ensure that they look and behave exactly the same on each platform is completely beyond the means of any library or archive in the world. This is quite aside from any legal restrictions which might apply to the activity.

A second approach to digital preservation moves attention away from digital objects and concentrates instead on the hardware platforms on which they were originally intended to run – their environments. The so-called 'emulation' approach attempts to develop software that when run on one hardware platform makes it behave as if it were another. Thus, one might run a "PC Emulator" on an iMac, thereby enabling one to access PC files on Apple hardware. The principal advantages of emulation are that, if done properly, it completely bypasses any concerns about file-format inflation and complexity. If the emulator performs as it should, then any file which ran on the original platform whatever its format or complexity should perform under emulation exactly as it did on the original.

For most people emulation as a preservation approach is poorly understood to the point of being technologically intimidating, and it is not uncommon even within the Digital Preservation community for mention of emulation to be met with incomprehension where it is not actually resisted. The complexities of producing an emulator are certainly great, and within a preservation context it is necessary not only to capture all the officially documented characteristics of a hardware system, but also to incorporate undocumented aspects of the original platform. This extra requirement is not only prompted by some sense of intellectual completeness but reflects the fact that a great deal of high performance software depends on exploiting (deliberately or otherwise) undocumented hardware features. The technical understanding required to produce such a completely faithful emulator is not widely available and in the case of some machines may not any longer exist. Therefore, while an emulation-based approach to digital preservation has definite attractions, it cannot be considered an easy alternative. Furthermore, emulators are digital objects in their own right, and are just as subject to being rendered inaccessible when the hardware platform on which they were designed to run becomes obsolete, as are any other digital objects.

As new hybrid disciplines, such as the Digital Humanities, start to emerge and mature, and as information technology enables researchers to conduct their work in virtual environments, the need to understand what is meant by 'preservation', what constitutes a 'digital object' and how digital research creates artefacts in their own right is becoming ever more critical. There are many myths, assumptions and implementation failures surrounding digital preservation and this chapter will examine some of these.

Digital objects and the data with which they are entwined have within a generation become not only commonplace but ubiquitous. Once the domain of specialist equipment, skills and tasks, the silicon revolution now gives anyone with access to a computing device the ability to access, create, manipulate and comment upon data. It can be argued that this familiarity, while not breeding contempt, has somewhat devalued our perception of data. While Moore's Law predicts a doubling of processing power every two years, Rock's Law (also known as Moore's Second Law) observes that the cost of producing the components doubles every four years, reflecting the expanding market for new and innovative products requiring capital investment which in turn perpetuates new products. This cycle is ultimately dependant on the marketing of new product to the masses and a high obsolescence attrition rate, especially when combined with software that takes advantage of new innovations and the repurposing of earlier developments into other ancillary hardware.

If tangible objects from computer workstations to smart phones, monitors to mass storage devices, have a relatively fixed commercial lifespan and are considered almost as consumable goods often not considered as items requiring accounted depreciation (at least in terms of components), then what hope is there for intangible assets such as data? The loss of an MP3 player may be inconvenient but a new and better model can be purchased at a similar, if not better price, than that paid for the older model, and the albums downloaded from the Internet. To replace a sound system that would allow one to play a vinyl collection, or even cassette tape, would cost significantly more.

If there is a commercial imperative then to sell product to the mass market there is an implied promise that the newer, faster, better, easier version will provide some form of integration with existing systems. While clearly if there is a radical change, for example for video tape to compact disc, one cannot expect the media to work (although one is always surprised to hear that someone cannot watch their wedding video anymore as they disposed of their VHS recorder some years ago), anyone faced with connecting for example a flat screen LED television to existing DVD players, game consoles, media stations, online services and/or home computer networks soon finds themselves lost in the different combination of cables and input/output options.

To some extent this naivety extends to the digital domain; we assume that someone is "looking after" the Internet and we are upset when our favourite website or service is taken down or unavailable, a DVD fails to play on our new system, a CD-ROM supposed to be able to store data for 100 years is corrupted or the latest update to a piece of software changes the way we work. And yet at the heart of this complex system of disposable parts lies the very thing which is set up to allow us to access - the data, the ghost in the machine, without which the whole system would be useless. If any part of this complex digital object fails then the integrity of the whole cannot be guaranteed: preservation of complex digital objects then is much more than ensuring that your files are backed up.

The Oxford English Dictionary defines preservation as “the act of maintaining (something) in its original or existing state”; this implies that the object of preservation is both stable and that there is a process of ensuring that its stability is continued. We do not commonly think of ‘digital’ as implying statis or stability; the idea that digital objects are dynamic, interactive, customisable and available on demand has more to do with good marketing for products than describing the actual data that it uses in novel ways. Data, when taken as a whole, is for the most part, static.

So if data is primarily static or at least if it can be captured at a point in time, a snapshot if you will, then the data could be hived off and kept somewhere safe ready for when it might be needed. While this is to some extent true, such a viewpoint focuses on the data, and as such data on its own is not of much use: it needs to be organised in some way, stored in some format and accessible by some method in order to be useful rather than just a series of bits and bytes. These additional properties when bundled with the data denote a data object (or record or file).

Much like a series of characters organised together as a piece of text, these bundles make sense of the words that they form and organise them into a comprehensible communication. This piece of text however is only comprehensible if it obeys the rules of the language it is written in and if the person reading it can understand that language. Our conceptual piece of text therefore must have some standards applied to it in order to be understood and the mechanism for reading the text must be available in order for the text to be actually read.

It is remarkable that so much tacit or assumed understanding of how data is organised and why formats and standards are important, is unacknowledged or undocumented in many digital projects. Whether this is a direct result of the ease of access and usage of information technology, or whether the drive to use information technology in new and innovative ways means that such considerations are sidelined, is open to debate.

Wherever the problem may lay, the greatest criticism of research utilising visualisation and simulation as part of its methodology is that it is not ‘scientific’, or rather that it does not follow the principles of scientific methodology and in particular that results cannot be readily reproduced. It does not mean that the data used to create the visualisation is necessarily flawed, rather that the results are different because the environment in which the data exists has changed. If, then, we can ensure that not only is the data, the characters of our text, are recorded but also the entire digital object is described, to extend the metaphor, how the book containing the text is organised, the conventions it uses and its dependencies, then the meaning of the text can be understood within the context of its environment of the book. The object book can be preserved in a library, reprinted, translated, scanned from paper to a data file or even emulated as an e-book but the knowledge that is created through the interaction between the data and its object is maintained, preserved and repeatable. Understanding this complexity and being able to communicate this with the data therefore is fundamental to preservation. It is this symbiotic relationship between data and the digital object and how that complexity can be managed, documented and safeguarded in a project that we will now consider in practical ways.

The first step in planning the preservation of a digital object (of whatever complexity) is to apply a known standard to the component parts of that object. As an object may have many different components it may be necessary to apply different standards to each different part, and it may be necessary to apply multiple standards to a specific media type depending on the circumstances. While the application of standards is crucial to the preservation process it is equally as important to record the standard that is being used as part of the metadata record of the object. The application of standards does not guarantee that digital object preservation is assured, but it greatly enhances the probability that the data will be comprehensible and useable long term.

Where proprietary formats cannot be avoided, then the preservation plan should strongly consider the use of a parallel standard that can be included within the preservation process as an ancillary archive item. A number of projects conducted by King's College London's King's Visualisation Lab utilised the online persistent virtual world platform of Linden Lab's 'Second Life®'.[1] While the platform offers the user the functionality to construct digital artefacts natively within the Second Life environment, the decision was taken to conduct the modelling phase of the project offline wherever possible using 3D Studio Max.

The 3D Studio Max file created to describe the geometry and the associated texture files (mostly held in Adobe's Photoshop .PSD format as working files) were converted into a form that could be imported into the online virtual world environment. In addition to the .max and transfer format the files describing the geometry were also exported into VRML97 ISO/IEC 14772-1:1997[2] format for archiving along with the original source files. The decision to build offline was primarily informed by the absence from the Second Life platform of any mechanism for saving or export from the virtual world (the environment is persistent) or provision of version control for development (and recovery). In addition there was an identified risk in the nature of the platform; Linden Lab is a commercial company and charges for renting virtual land and could change the terms and conditions of usage or close down thus rendering the work inaccessible, and because of the persistent nature of the development of the platform, the system is, in effect, in a state of perpetual beta testing with no single point that could be selected for a base line of preservation.

At first glance the process of producing a 'data double' may seem superfluous but such redundancy can be seen as an extension to the quality control process in ensuring that a build has interoperability between platforms as well as producing a data recovery strategy as part of a project's overall security procedures. For those in the visualisation community this concept will not be new; the creation of independent digital objects or data artefacts is common best practice. Each component will have the geometry, properties and behaviour for each artefact as well as any metadata and paradata stored as an integral part of the data load for that artefact. The construction of a data artefact is in itself an act of preservation and while it may seem time-consuming, it does not necessarily have to be especially so if the artefact is intended to be reused, software engineering principles are applied and the appropriate resources have been allocated to the development cycle for the project.

The starting point for creating any data artefact is the specification document which will describe the object, its behaviour, requirements and any subordinate objects. From this all geometry, code, component parts, metadata and paradata will be derived. When

[1] http://secondlife.com/

[2] ISO/IEC 14772-1:1997 Information technology—Computer graphics and image processing—The Virtual Reality Modeling Language—Part 1: Functional specification and UTF-8 encoding. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=25508

the inclusion of a data double is required then it should be included in the specification document and its relationship shown in the entity relationship diagram. Whether the data double should itself reference other standardised data double types or not is a matter of preference but should be clear from the documentation. If the data double has graceful failure functionality, such as looking for a resource in multiple locations via URL, then this can be used both to add robustness to the integrity of the object and identify the standard resource.

Documentation is the scourge of many projects but is vital to communicate purpose, structure and process between the different team members and end users, which includes those who will be involved with the maintenance of digital objects. Again just having good documentation will not preserve a visualisation or even a digital object but it will help to ensure that the concepts and direction are detailed and provide a record of what was supposed to be included even if the digital object integrity has failed. It should be noted that inline comments within source code and house style are also documentation in their own right and care should be taken when finalising code to ensure that unnecessary comments are removed and that house styles are applied throughout.

Good documentation and data artefact creation can only be achieved when there is an understanding of the benefits of doing so and where there is an institutional willingness to invest the time and resource in their creation. It is estimated that correct application of paradata and data artefact creation will add up to 50% to the total planned time of the development cycle, and further that this must be planned as an integrated activity of development. Regrettably additional development time for this process is rarely included in the original project schedule or considered an additional process that can be slotted in at the end of the project. As project time tables get squeezed, clients change specifications, time is spent developing feature creep functionality or it simply happens that the development cycle has been under-resourced; the paradata process is abandoned in favour of meeting delivery schedules.

Maintaining good documentation and adhering to standards may not be very exciting, especially to those who are used to working in visual media with almost real time results. However, failing to do so will cause long term problems for preservation. It is not difficult to find examples of projects where noncompliance with standards and documentation have caused significant problems either when maintenance was required, or archived work was revived for inclusion in later projects. In one case, the source code for an entire content management system written by a commercial external partner that closed down at the end of the project, was handed over in a mixture of English and Dutch, with no visible conventions and no documentation; eventually the entire backend had to be rewritten.

One possible conclusion that can be drawn from the previous example is that it was not the absence of standards and documentation that caused the failure of the system and subsequent work, but rather that a change in a third party piece of software required alterations in the system which, if properly documented, would have been relatively straightforward to implement. When considering preservation it is all too easy to focus only on just the data or the bespoke software that has been created as part of a project and make the assumption that everything else will remain the same or at least be compatible. If we recall the definition of preservation, the phrase ‘original or existing state’ is a useful reminder to include provision for ancillary software and hardware on which the system is dependent.

To misquote Benjamin Franklin: “In this world nothing can be said to be certain, except death and taxes and upgrades”. Generally upgrades are considered a good thing, often bringing new functionality and productivity, but when seen through the lens of preservation these benefits are not relevant as the systems and data objects have been built to utilise known functionality at the time of creation, and rarely have over-specified affordances that offer even medium-term future-proofing.

In some projects it is the case that although standards are applied to data objects and their properties and interconnections understood and described, the systems that supported them are external to the project. Of these one of the longest running, THEATRON, an online e-learning suite for theatre history, was constructed around HTML and VRML data objects enhanced by a variety of rich media. The output of the project was launched to the general public in 1999 to both critical and popular acclaim and still attracts around 300 unique users each month. Over a decade after its launch, the system has suffered numerous problems with changes in third party software, and web browsers have changed the way that they display information, perform operations and allow access to resources, VRML has not achieved the predicted uptake for 3D display on the Internet, component software has been deleted from catalogues restricting availability and functionality and codecs and media players have changed rendering movies unplayable.

Not only must developers consider and prepare for how they will ensure their data is fit for archiving, but they must also preserve how that data will be accessed. The digital objects and systems to run THEATRON exist in a 'perfect' archive state as the project still holds a fallback server which was created at time of publication. As such, this server is a 'closed' digital object, and apart from hardware failure and data loss could be seen by some as representative of digital preservation. However this perfect state archive of the digital object is relatively useless, apart from providing a reference point should that project ever be revisited (simply providing a data 'quarry' for a reworking of the application) as no client system required to connect to and access the held data has been archived along with the server.

Even if a system is considered to be closed and not requiring additional software there will be, in almost every case, software components outside the control of the developer. For example, operating systems, device drivers, network and communications software (not to mention lower level firmware) all form part of the electronic ecosystem within which digital objects reside. Preserving all of this software may not be viable as part of the preservation process (see the chapter of D. Anderson) but as a minimum the versions of all software used by the final system should be documented, giving name, version and release and, if appropriate, any additional information such as customised settings.

If the digital object is utilising external objects from the Internet, such as web page links or resources then, once again, we must consider that these are volatile and may not either be available or contain the same information as originally held. One project contained links to websites from which useful tools had been acquired only to discover some time later that many of these were no longer available and had been removed by the provider, and in one case the domain name had expired and been bought and reissued as an adult entertainment site. Wherever possible linking to online web resources should take full advantage of alternative location descriptors and, it is recommended that within the documentation all web-linked resources carry a citation similar to those used when authoring written papers giving a retrieval date of known accessibility.

Similarly the preservation of systems may be compromised by changes in hardware technologies which cannot be predicted. For the visualisation community these are arguably the graphics processors and display technologies. While graphics cards usually provide some basic functionality and backwards compatibility, this cannot be assumed especially where specific affordance of hardware is being exploited. The recent uptake of LCD and LED screens over CRT has implications too for preservation of the integrity of the visual image. Screen resolutions have changed as have refresh rates, and applications designed with older resolutions in mind may find unexpected results such as image distortion when run though new displays. Where image quality is important, for instance in lighting simulations, calibration may no longer be possible or may provide less than optimum results.

For both hardware and software the minimum specifications for the system will be provided by the project's requirement documents, which can be supplemented by the quality assurance testing reports prior to release. While such documentation will not protect the system from the progress of technology, it will ensure that future investigators can understand the base requirements of the system and the range of functionality that was available over a range of different test combinations. All information of a technical nature will help in understanding what functionality is required in reconstructing the system either physically or through emulation.

It should go without saying that specialist hardware built specifically for or utilising features of specific hardware components will not necessarily work or be able to use the intended functionality of newer hardware. KVL had experience of this in a proof of concept pilot project using early consumer market 3D shutter glasses. The system required to run the proof of concept application depended on a specific combination of graphics card, drivers and a high refresh rate CRT monitor. Changes to any of these individual components resulted in unexpected results and failure, most notably the drivers which contained an exploit used by the shutter glasses was reissued closing out the functionality and requiring the use of older additional software that in turn excluded affordances to other, newer software. In all of this the data component of the digital object still exists, uncompromised by the failure of the hardware component but is useless as it is bound to the hardware environment.

The third part of the system trinity is the human component, sometimes referred to as "wetware", and constitutes the knowledge of how to operate a system of hardware and software. Preserving that knowledge and those skill sets is probably the hardest of all of the components of the system; users tend to learn by building on tacit knowledge of systems which retain familiarity over generations of software/hardware, but seldom consider reading the operation manual for the latest piece of equipment unless they need specific information about new or changed functionality. However, such a continuity of understanding cannot be assumed when looking to the future. We have already touched on the importance of documentation for systems specification and functional requirements and as a minimum these should be incorporated into the digital object. Furthermore any operation documentation for the system and, if appropriate, links to online support or community-derived resources should be included.

Having looked at ways in which the developer and project can prepare complex digital objects for a process of preservation, or at least understand and document the boundaries of preservable items, we must consider the implications of the preservation process. This volume discusses many different approaches to the practicalities of preservation, from best practice through to third party hosting and archiving which gives a project the opportunity to consider and lay down plans at the start of the development process. It further contains links to organisations and best practice resources that will help guide the reader in ensuring that their complex digital objects have the best chance of being preserved in a state which will be suitable for future generations to access, use and build upon.

There is however one critical obstacle that must be overcome if preservation is to occur and that is the institutional support and vision. There appears to be a perception that preservation of digital objects is simply achieved by ensuring that data can be placed onto storage media, put on the shelves of an archive and, much like physical books, will be there in perpetuity for all. We have seen that complex digital objects are much more than simply the data; if it were just that then the solution to preservation would be as simple as printing the data out onto paper and archiving it applying the knowledge, techniques and wisdom of the archive curator and librarian who have centuries of understanding the medium and how best to manage it. However, the complexity that arises from the interaction between data, software and hardware is much more subtle and ephemeral than crude methods (with apologies to archivists and librarians everywhere) and with such a dynamic and volatile set of variables a traditional, and often intuitional, view of preservation cannot apply to the digital domain.

If the institution is not prepared to accept the value of digital objects, it will not accept the cost and effort in preserving them. Digital is not disposable anymore; as we start the migration from the physical to the virtual experience and experimentation, both the intellectual capital and cash price invested in producing digital objects is increasing. With the ever-diminishing pot of funding both institutionally and from sponsors, provision is seldom made for preservation post completion of a project or, if there is, at a fixed term. If such a commitment is not forthcoming, then the preservation of complex digital objects will rely on the level of care that the developers have built into the object in the hope that some day someone will be able to resurrect it.

There are alternative possibilities that could be considered. Funding could be sought explicitly for preservation purposes; although this is unlikely to cover everything, it may preserve specific components. Funding to extend the initial project or refresh digital objects to the latest benchmarks is possible but traps the project in a constant recycling of material that does little more than explore the functionality and affordances of new technical developments unless new research questions and directions can be developed from the existing objects.

As a discipline, computer science has been in existence for around sixty years, widely accessible personal computing for thirty and the Internet/world wide web as we recognise it today less than twenty. Digital preservation is a young but vital field which demands that we understand what is possible in the present and at the same time plan and predict for the future. We must practice *Ruinenwerttheorie*, the theory of ruin value, building digital monuments today while trying to visualise their future use and relevance to posterity.

Current Digital Preservation Approaches and Practices in Simulation and Visualisation Projects

Digital Preservation and Curation: The Danger of Overlooking Software

Neil Chue Hong

Software Sustainability Institute, The University of Edinburgh, James Clerk Maxwell Building, Mayfield Road Edinburgh EH9 3JZ

Abstract. Software is often overlooked when considering the wider requirements for preserving digital outputs. A key challenge in digital preservation is being able to articulate, and ideally prove, the need for preservation. The Software Sustainability Institute in partnership with Curtis+Cartwright Consulting have published a series of outputs to support the sector by raising awareness of software sustainability and preservation issues. This chapter summarises some of the advice set out in the Software Preservation Benefits Framework that has been developed which can help groups understand and gauge the benefits or drawbacks of allocating effort to ensuring that preservation measures are built into processes and to promote actively preserving legacy software. This identifies four key purposes, and seven different approaches, to preserving software. In this way, important software can be preserved for future generations.

Introduction

From preserving research results, to storing photos for the benefit of future generations, the importance of preserving digital data is gaining widespread acceptance. But what about software?

It is easy to focus on the preservation of data and other digital objects, like images and music samples, because they are generally seen as end products. The software that is needed to access the preserved data is frequently overlooked in the preservation process. But without the right software, it could be impossible to access the preserved data – which undermines the reason for storing the data in the first place.

A key challenge in digital preservation is being able to articulate, and ideally prove, the need for preservation. There are different purposes and benefits which facilitate making the case for preservation. These should be combined with preservation plans regarding data and hardware: digital preservation should be considered in an integrated manner. For example, media obsolescence and recovery is often as much a part of a software preservation project as a data preservation project.

The Software Sustainability Institute[1] in partnership with Curtis+Cartwright Consulting[2] have developed a series of outputs to support the sector by raising awareness of software sustainability and preservation issues, as part of a JISC-funded initiative. In particular a Benefits Framework[3] has been published that can help groups understand and gauge the benefits or drawbacks of allocating effort to ensuring that preservation measures are built into processes, and to promote actively preserving legacy software. The rest of this chapter provides a summary of advice based on this work.

[1] The Software Sustainability Institute: http://www.software.ac.uk/

[2] Curtis+Cartwright Consulting: http://www.curtiscartwright.co.uk/

[3] Benefits Framework: http://www.software.ac.uk/attach/SoftwarePreservationBenefitsFramework.pdf

When should you Consider Software Preservation?

Software is used to create, interpret, present, manipulate and manage data. There is no simple and universally applicable formula for determining if your software needs to be preserved, and how to go about preserving it. Instead there are a range of questions and factors which should be taken into account. In particular, curators should consider software preservation whenever one or more of the following statements is/are true:

1. The Software cannot be Separated from the Data or Digital Object

In an ideal world, data can be isolated and preserved independently of the software used to create or access it. Sometimes this is not possible. For example, if the software and the data form an integrated model, the data by itself is meaningless. This means that the software must be preserved with the data to ensure continued access.

If data is stored in a format that is open and human-readable, then any software that follows that format can be used to read the data. If the data is stored in a format that is closed and arcane, then you must also preserve the software that is used to access it.

2. The Software is Classified as a Research Output

The software could fall under a research funders' preservation policy. This means that the software must be preserved as a condition of its funding. It may also be the case that software must be preserved to achieve legal compliance.

3. The Software has Intrinsic Value

Software can be a valuable historical resource. If the software was the first example of its type, or it was a fundamental part of a historically significant event, then the software has inherent heritage value and should be preserved.

Challenges of Preserving Software

Software presents additional challenges to those who curate, preserve and archive digital artefacts. In particular, software preservation is difficult because software is sensitive to changes in its environment.

If there is a change to the computer or operating system on which the software runs, the software will often stop working properly. What is more, this change might not cause a catastrophic failure. Although serious, this kind of failure is at least easy to spot. A change to the computer or operating system might only cause a subtle, yet important, change in results. Expert knowledge is needed to fully understand how a software component works and the effect that a change may have. This is particularly the case where it may not be easy or possible to recreate the original conditions or input data, for example with simulations of nuclear explosions.

However software preservation is not the same as software engineering. Although there is a wide overlap between good practice in each, software engineering is mainly concerned with maintaining, fixing and increasing the functionality of current systems, whilst software preservation often focuses on maintaining reproducibility of past performance. In general, good software preservation arises from good software engineering, and software engineering best practice such as clear licensing; clear documentation; use of commonly adopted and modern programming languages; modular design; revision management and change control; established software testing regime and validated results; separation between data and code; and clear understanding of dependencies all make the work of software preservation easier.

There is a lot of variation in software: it comes in many different forms, it is written in a bewildering range of languages and it can be licensed in many different ways. Further difficulties can arise from the increasing use of web services and the cloud. This is where your software is hosted by external organisations – a practice that is becoming increasingly popular. Choosing what to preserve can be difficult. As software becomes increasingly transient, it may be that the workflow or parameters become more important than the code itself. Especially where the software is an enabler, rather than something that must be preserved in its own right, it is advisable to try and recast a software preservation problem into a data preservation one as they are invariably easier to handle.

This means that it is important to understand all the different facets of software and choose the best route for its preservation. Software preservation should be part of a broader preservation strategy. This strategy should provide a guide of what needs to be preserved, and for how long.

How should I Approach Software Preservation?

When considering software preservation, you should consider the following questions:

- Is there still knowledge and expertise to handle and run the software?
- How much access do you have to the:
    - owners;
    - developers;
    - source code;
    - hardware;
    - users?
- Do you have the necessary Intellectual Property Rights (IPR)? (See the chapter on The Impact of European Copyright Legislation on Digital Preservation Activity: Lessons learned from Legal Studies commissioned by the KEEP project by David Anderson in this book).
- How authentic does the preserved software need to be?
    - exactly as the original;
    - as the original but fixing errors when found
    - mostly the same but tolerant of minor deviations; or
    - only replicating certain pieces of functionality?
- What is the maintainability of underlying hardware?
- Is maintaining integrity and/or authenticity an important requirement?
- Is the software covered by a preservation policy / strategy?
- Is there a clear purpose in preserving the software?
- Is there a clear time period for preservation?

- Do the predicted benefit(s) exceed the predicted cost(s)?
- Is there motivation for preserving the software?
- Are you also interested in further development or maintenance?
- Is the necessary capability available?
- Is the necessary capacity available?
- Are the necessary resources available, and at the right times?

The same considerations that apply to digital preservation also apply to software (intellectual property, choice of media, backup and recovery, etc.), so the basic considerations of software preservation are similar to those of digital preservation.

Preserving the knowledge behind software is as critical as the software itself. Good documentation is important, as is having access to the developers of the software. A project undertaken by the STFC identified a set of significant properties of software[1], which can be used as a structured framework to elicit key information from the development team.

Purposes and Benefits of Software Preservation

A key challenge in digital preservation is being able to articulate, and ideally prove, the need for preservation. The benefits framework developed by the SSI and Curtis + Cartwright identifies four main purposes to software preservation, along with their associated benefits. A summary of these benefits and a range of illustrative scenarios for each purpose are listed in the following table:

Table 1. Benefits and scenarios

Purpose	Benefits	Scenarios
Achieve legal compliance and accountability	Reduced exposure to legal risks Avoidance of liability actions Easily demonstrable compliance lessens audit burden Improved institutional governance Enhanced reputation	Maintaining records or audit trail Demonstrating integrity and authenticity of data and systems Addressing specific contractual requirements Addressing specific regulatory requirements Resolving copyright or patent disputes Addressing the need to revert back to earlier versions due to IP settlements Publishing research openly for transparency Publishing research openly as a condition of funding
Create heritage value	(Heritage value is generally considered to be of intrinsic value)	Ensuring a complete record of research outputs where software is an intermediate or final output Preserving computing capabilities (software with or without hardware) that is considered to have intrinsic value Supporting the work of museums and archives
Enable continued access to data and services	For research data and business intelligence: • Fewer unintentional errors due to increased scrutiny • Reduced deliberate research fraud	Reproducing and verifying research results Repeating and verifying research results (using the same or similar setup) Reanalysing data in the light of new theories Reusing data in combination with future data

[1] http://www.jisc.ac.uk/media/documents/programmes/preservation/spsoftware_report_redacted.pdf

Encourage software reuse	Reduced development cost Reduced development risk Accelerated development Increased quality and dependability Focused use of specialists Standards compliance Reduced duplication Learning from others Opportunities for commercialisation	Continuing operational use in institution Increasing uptake elsewhere Promoting good software

Seven Approaches to Software Preservation

As part of the study, we observed seven distinct approaches, characterised by what is being preserved and passed on. The approaches set out are an extended set from those traditionally covered. This is done to give proper coverage of approaches specific to open source software, to include a more pragmatic option that gives additional flexibility and to include a 'default' option for a baseline. The different approaches therefore covered are:

- Technical preservation (techno-centric) – Preserve original hardware and software in same state;
- Emulation (data-centric) – Emulate original hardware / operating environment, keeping software in same state;
- Migration (functionality-centric) – Update software as required to maintain same functionality, porting/transferring before platform obsolescence;
- Cultivation (process-centric) – Keep software 'alive' by moving to a more open development model, bringing on board additional contributors and spreading knowledge of process;
- Hibernation (knowledge-centric) – Preserve the knowledge of how to resuscitate/recreate the exact functionality of the software at a later date;
- Deprecation – Formally retire the software without leaving the option of resuscitation/recreation;
- Procrastination – Do nothing.

Some approaches are better suited than others to each purpose. The table below provides an indicative mapping between the four purposes and appropriate approaches. We do not consider procrastination to be an appropriate approach to any software.

Each approach is provided with a description, a set of activities, and notes on costs. Some metrics (or more general indicators) are also proposed to determine if the approach is going to plan. These will need to be tailored to the specific software preservation plan being used. A suitable set of metrics will inevitably cover a broader scope than software preservation and include technical, process and economic factors.

Table 2. Preservation approaches

	Technical Preservati on	**Emulation**	**Migration**	**Cultivati on**	**Hiberna tion**
Achieve legal compliance and accountability	✓	✓	✓		
Create heritage	✓	✓			

value					
Enable continued access to data and services	✓	✓	✓	✓	✓
Encourage software reuse			✓	✓	✓

Technical preservation

Technical preservation is a planned and intentional decision to keep the software and hardware running in the same state. There is also the option of purchasing spares so that components can be replaced as they fail. It is important to bear in mind that no obsolete technology can be kept functional indefinitely. Technical preservation generally works best for when there is a known preservation period (especially if this is less than known support periods).

Software is reliant on hardware, and hardware changes as each new model is released. Over time, hardware will change to such an extent that older software will not run on the latest hardware. Without the hardware to run on, software becomes redundant.

The easiest way to ensure that there will always be hardware to run your software is to preserve the hardware. Technical preservation has one big benefit: it is easy in principle: you simply continue business as usual. However, there are drawbacks to this approach, the first being maintenance. Over time, hardware components will wear out and must be replaced. If the hardware is no longer manufactured, components become scarce and expensive. Ultimately, you may find yourself with broken hardware and no way of fixing it – leaving you with redundant software. The second drawback is isolation. Your software only works with very specific hardware, which limits your users to those people with the right hardware. This might be a very small group. Technical preservation is a straightforward approach to sustainability, but it is only as reliable whilst you have a stockpile of spare parts.

Specific activities within this approach are likely to include:

- purchasing spares;
- regular checking that the system works;
- maintaining hardware;
- replacing hardware elements as they fail;
- scheduling review points in the calendar.

There are some points to factor in about the costs of this approach:

- Some upfront costs to purchase spares;
- Low cost initially (maintenance only) to keep the hardware and software running;
- Costs likely to rise over time as maintenance gradually becomes more difficult;
- At some point a large cost will be incurred as hardware fails and a replacement approach is necessary.

Some metrics or indicators to monitor to know how well this approach is proceeding could be designed around one or more of the following:

- continued executability (e.g. percentage monthly uptime, or number of failures a month);
- ongoing maintenance overheads (e.g. effort per month, or direct/indirect cost per month);
- number of remaining spares;

- expected cost of a replacement system (NB: this will change non-linearly over time).

Emulation

You want to keep your software, but you are worried that technical-preservation might leave you with no hardware or an expensive maintenance bill. The alternative may be emulation. An emulator is a software package that mimics your old hardware and/or operating environment, and can be run on any computer.

Emulation gives you the flexibility to run your software on new hardware, which gives your software a new lease of life. As always, there are drawbacks. You need to find an emulator. You might be lucky and find one available under a free-to-use licence, or you might be able to buy one. However, if your old hardware was rare, you may find that no emulator exists. In this case, you either have to write an emulator yourself, which requires specialist skills and could be expensive, or explore another of the sustainability approaches. It is difficult to write an emulator that perfectly mimics the old hardware. This can lead to differences between the operation of the old hardware and the new emulator, which could manifest themselves in annoying quirks or more serious problems.

Specific activities within this approach are likely to include:

- regular checking that the system works;
- regression testing;
- verifying and validating results;
- updating the emulator (or maintaining it if developed in-house);
- scheduling review points in the calendar.

There are some points to factor in about the costs of this approach:

- Low cost if emulator exists as costs borne by someone else;
- Emulators themselves need sustaining;
- At some point a large cost may be incurred as emulator ceases to work and a replacement approach is necessary.

Some metrics or indicators to monitor to know how well this approach is proceeding could be designed around one or more of the following:

- continued executability (e.g. percentage monthly uptime, or number of failures a month);
- frequency of updates to emulator to know if the emulator is being sustained (e.g. updates per quarter, number of unresolved bugs);
- cost of emulation (e.g. total costs in licensing, handling emulation errors, verifying data, etc. per month or quarter);
- emulation performance (e.g. average response time over a month for a service or the program execution time for a command-line application based on some sequence of test queries or input).

Cultivation

Sustainability requires the investment of resources, and cultivation is one of the best ways of sharing the responsibility for these resources. Cultivation is the process of opening development of your software. This is where you allow developers access to your code – under a licence – so that they can work with you. The deal is that outside developers can develop your software so that it meets their exact needs, and in doing so, any bugs they fix or new functionality they add can be given back to your project.

Cultivation allows more contributors to be brought into your project, which helps share the sustainability workload. With more people, knowledge about your software is spread over a wider group, so that the departure of one person is less likely to affect the software's future.

The main drawback of cultivation is that it is a long-term process. Cultivation is not suitable as a quick fix to ensure sustainability in the short term; instead it requires effort and planning over many months and years. Moving to open development is not as simple as making your source code publicly available. You also need to build a community around the software, and this requires work to understand your community and how to appeal to them. Once in place, your community could become self-sustaining so that the future of your software is assured.

Cultivation promises a self-sustaining community of developers who work together to keep your software up to date, but requires work to cultivate the right community for your software. A combination of Open Source licensing and Open Development practices make it easier to preserve software by removing barriers to others taking on the preservation of the code. The body of knowledge about a piece of software is more likely to be manifested in electronic form, as opposed to being held in the heads of a few developers. However it is important to reiterate that Open Source Software (OSS) alone is not enough to enable the preservation of software – this also requires aspects of curation and ongoing minimal maintenance to cope with environmental changes.

Specific activities within this approach are likely to include:

- choosing an appropriate open source licence;
- applying an open source licence to an existing codebase;
- moving code to an open source repository (e.g. sourceforge.net);
- setting up a development website, mailing list, etc.;
- cleaning code to make it presentable for new comers;
- providing test data for everyone to use to validate functionality;
- establishing governance for the software;
- engaging with users and contributors and listening to feedback and ideas;
- scheduling review points in the calendar.

There are some points to factor in about the costs of this approach:

- Ensuring a sufficient maturity of software could add significant costs (e.g. taking a prototype to reusable software is often a factor of 10);
- Cultivation will involve a sustained effort to move to a more open development model;
- The costs and likelihood of eventual success are difficult to predict;
- If it is successful, it will spread costs over a larger number of individuals and organisations;
- The ideal outcome is that it becomes financially self-sustaining.

Some metrics or indicators to monitor to know how well this approach is proceeding could be designed around one or more of the following:

- OSS Watch Software Sustainability Maturity Model (i.e. the SSMM Level)[1];
- the Community Roundtable's Community Maturity Model[2] (i.e. which stage from Hierarchy to Emergent Community to Community to Network);

[1] http://www.oss-watch.ac.uk/resources/ssmm.xml

[2] http://community-roundtable.com/2009/06/the-community-maturity-model/

- size of user community: rocketing / increasing / a ‘known’ community / decreasing / plummeting (e.g. specifically the number of individual active users);
- spread of user community: internal / external / cross-domain;
- number and spread of contributors (e.g. number of individual contributors, or the number of contributions from specific communities);
- continued executability (e.g. percentage monthly uptime, or number of failures a month);
- continued compilability (e.g. binary yes/no for compilation failure, or number of compilation warnings).

Hibernation

Rather than sustaining your software as operational software, you may choose to hibernate it. You may choose hibernation when your software has come to the end of its useful life, but may need to resurrect it to double-check analysis or prove a result. Alternatively, there may not be a user community for your software, but you believe one will occur in the future. Hibernation allows you to preserve the knowledge about your software so that it can be resurrected in the future.

Hibernation can be a one-off process. Unlike sustainability, which requires a continuous investment of resources, the hibernation process can have a beginning and – importantly – an end. Preparing software for hibernation can be resource heavy, and if the software is never resurrected, you may feel that those resources were wasted. Hibernation allows you to store software that you do not currently need, but it requires a significant – if short lived – investment of resources.

Specific activities within this approach are likely to include:

- reviewing and improving documentation;
- recording the significant properties of software;
- archiving the software along with all documentation;
- scheduling review points in the calendar.

If the software is already OSS then hibernation should be relatively straightforward, since there ought to be a code repository, up to date documentation and a means to contact user and contributors (if any).

There are some points to factor in about the costs of this approach:

- At its simplest, hibernation involves documenting pseudo-code (e.g. publishing the algorithm in a research paper) – this is inexpensive;
- However, ensuring rigorous documentation, test data, etc. is time consuming;
- There is a small on-going cost to ensure discoverability, accessibility, etc. of hibernated software and materials;
- The big advantage of hibernation is that it should significantly reduce future development costs.

Some metrics or indicators to monitor to know how well this approach is proceeding could be designed around one or more of the following:

- completeness of documentation (code, design, testing, etc) (e.g. percentage completion of the Significant Properties Framework[1]);

[1] http://escholarship.org/uc/item/8089m1v1.pdf

- currency of programming language, middleware and operating environment (e.g. number of updates in the past year, number of updates planned for next year, time to end-of-support-life in months, total number of months past end-of-support-life, or total number of major new releases that supersede the version used);
- archive availability and resilience (e.g. that defined by Service Level Agreements, frequency of backup, or percentage up time in the last month);
- compilability and executability at review points (e.g. see those defined in Migration, above).

Deprecation

If software lacks a community, the resources to continue or a developer, then the only alternative is deprecation. All effort invested into the software comes to an end, but, unlike hibernation, no effort is invested in preparing the software beforehand. In the future, if someone wants to use the software, they may not be able to find a stored copy and it might be expensive or impossible to resurrect the software.

Deprecation is easy to perform, but often marks the end of a software package's life and is typically only chosen when no other option is available. Deprecation is effectively enforced technical preservation – but without the thought and preparation to have any confidence that it will work as an effective preservation strategy.

If the software is then required, then you will need to provide, or buy in services for, software archaeology. This involves rescuing software from obsolete or damaged hardware, media and software environments – consider it an emergency recovery strategy. It may involve media recovery, for example if the media is heavily damaged, but more likely the real problem will be in understanding the code or binary. If you just have a binary then further information is probably necessary to determine what environment the software can be run on. If you have the code then at least you are able to adapt it to make it run, but the code may start off as being effectively unintelligible. It will be especially difficult to recover if the code is old, poorly documented, lacking the original build tools (compilers, makefiles, etc.). If it is not your code, or you have switched to using another programming language, then clearly it will be harder still. And without having pre-planned test data it will be hard to get the assurance that the software runs as intended – critical if you are after perfect repeatability.

Specific activities within this approach are likely to include:

- deciding on a timeframe for deprecation;
- notifying users and contributors of the intent to deprecate;
- archiving the software along with all documentation.

There are some points to factor in about the costs of this approach:

- There are short term costs in formally shutting down development;
- Depreciation generally assumes software has been superseded and no emergency recovery effort is needed.

Some metrics or indicators to monitor to know how well this approach is proceeding could be designed around one or more of the following:

- infrequency of user engagement;
- completeness of documentation (e.g. see those defined in Hibernation, above);
- archive availability and resilience (e.g. see those defined in Hibernation, above).

Procrastination

Procrastination is the default (but not recommended) option. It does not require any changes to the current working practices, and it does not involve any additional effort at the current time. However it can result in large amounts of effort needing to be expended in the future to continue to use the software, or end up wasting effort if the software is not required.

Building it into the Process

A principle from other areas of digital preservation is that considering preservation and sustainability upfront (and regularly) is important. This would imply that building preservation measures into software development and digital curation measures is good practice. Two 'preservation measures' are apparent:

- Software engineering: being able to encourage better software engineering practice early in the lifecycle will benefit software preservation if and when required.
- Identifying explicit preservation requirements: Requirements capture and management is an upfront activity in software development, and preservation requirements should be considered along with other requirements.

This suggests that it is important that there is a link between digital curators and software developers. The extent to which both software engineering practice and preservation requirements should be a priority depends on both the intended functionality of the software (i.e. whether it fits one or more of the four purposes) and the nature of the software itself. For example, is the software meant to be a proof-of-concept demonstrator, something more heavyweight like a pilot, or perhaps an operational service for a defined set of users? Each allows a different approach to be taken with different expectations of robustness and longevity.

It is important to understand all facets of software and choose the best route for its preservation as part of a broader preservation and development strategy. In this way, we can ensure that important software is preserved for future generations.

Acknowledgements. This work has been funded by the JISC as part of the Clarifying the Purposes and Benefits of Preserving Software study, carried out by Matt Shreeve (Curtis and Cartwright Consulting Ltd.) and Neil Chue Hong, Steve Crouch, Simon Hettrick and Tim Parkinson (Software Sustainability Institute). The Software Sustainability Institute is funded by the EPSRC under grant EP/H043160/1. Parts of this work build on the JISC-funded Significant Properties of Software study carried out by Brian Matthews *et al* at STFC.

Further Reading

Berman, F. et al. (2010). Sustainable economics for a digital planet: Ensuring long-term access to digital information. Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. 110 pp. Retrieved from http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

Chue Hong, N., Crouch, S., Hettrick, S., Parkinson, T., & Shreeve, M. (2010). Software Preservation Benefits Framework. 74 pp. Retrieved from http://www.software.ac.uk/attach/SoftwarePreservationBenefitsFramework.pdf

Matthews, B., McIlwrath, B., Giaretta, D., & Conway, E. (2008). The Significant Properties of Software: A Study, 99 pp. Retrieved from http://www.jisc.ac.uk/whatwedo/programmes/preservation/2008sigprops

How do I know that I have Preserved Software?

Brian Matthews, Arif Shaon, Esther Conway

e-Science Centre, STFC Rutherford Appleton Laboratory , Chilton, Didcot, OX11 0QX

Abstract. In this paper we consider some of the issues and barriers that have made the preservation of software appear dauntingly complex. We consider how different approaches can be made to preserve software and also the role of the supplementary information and digital objects which are needed to preserve software. We consider how we then might demonstrate that the preservation is actually successful, by discussing the notion of *adequacy* in software preservation, highlighting the need for the testing of software performance. We conclude with a discussion of tools and methods we are developing to introduce software preservation into preservation planning and support.

Introduction: Software Preservation

Software is a class of electronic object which is by its very nature digital and which is often a vital pre-requisite to the preservation of other electronic objects. However, software has many characteristics that make its preservation substantially more challenging than that of many other types of digital object. Software is inherently complex – forbiddingly so for people who were not involved in its development but nevertheless want to maintain access to software. A typical software artefact has a large number of components related in a dependency graph, and with specification, source and binary components, and a highly sensitive dependency on the operating environment. Handling this complexity is a major barrier to the preservation of software. Furthermore, the preservation of software is frequently seen as a secondary activity and one with limited usefulness. Software preservation is thus a relatively underexplored topic of research and there is little practical experience in the field of software preservation as such. In this paper, we lay out some issues that need to be considered in the preservation of software. Software can be defined as:

> *"a collection of computer programs and related data that provide the instructions for telling a computer what to do and how to do it. In other words, software is a conceptual entity which is a set of computer programs, procedures, and associated documentation concerned with the operation of a data processing system"*.[1]

Computer programs themselves are sequences of formal rules or instructions to a processor to enable it to execute a specific task or function. However, note that the definition also includes documentation, a crucial element in managing effective software preservation. We refer to a single collection of software artefacts that are brought together for an identifiable broad purpose as a software *product*.[2]

[1] http://en.wikipedia.org/wiki/Software. Retrieved 22 August 2011.

[2] Other alternative terms are also used, including software *system* (which could be confused with a complete assemblages of different hardware software items),and software *package* (which in the context of preservation, could be confused with the OAIS notion of *information package*). We chose the term *product* as a neutral term, and it should not imply that the software product is being provided on a commercial basis.

The term software is sometimes used in a broader context to describe any electronic media *content* which embodies expressions of ideas stored on film, tapes, records etc. for recall and replay by some (typically but not always) electronic device. For example, a piece of music stored for reproduction on vinyl disc or compact disc is sometimes described as the software for the record or CD player, in analogy to the instructions of a computer. However in this paper, such content is considered a data format for a different digital object type, and is thus out of scope.

Even considering this narrower focus, software remains a very large area with a huge variation in terms of nature and scale, with a spectrum including microcode, real-time control, operating systems, business systems, desktop applications, distributed systems, and expert systems, with an equally wide range of applications. There are also varying constraints of the business context in which the software is developed from systems coded by one individual for their own use (typical in research), open-source systems, to commercial products. We can classify this diversity along a number of different axes, which impact on preservation requirements.

- **Diversity of application.** Software is used in almost every domain of human activity**.** Thus there are software products in, for example, business office systems, scientific analysis applications, navigation systems, industrial control systems, electronic commerce, photography, art and music media systems. Each area has different functional characteristics on at least a conceptual user domain level and it is necessary to classify software according to some application-oriented classification or description of the domain.
- **Diversity in hardware architecture.** Software is designed to run on a large range of different computer configurations and architectures, and indeed "levels" of abstraction in relation to the raw electronics of the underlying computing hardware. At a micro level, assembler and micro-code are used to control the hardware directly and low-level operations such as memory management or drivers for hardware devices. At a higher level of abstraction, applications are intended to be deployed on a wide range of computing hardware and architectures (e.g. workstations, hand-held or mobile devices, main-frame computers, clusters). In order to recreate the functionality of system, the hardware configuration may need to be taken into account.
- **Diversity in software architecture.** Even within a common hardware configuration, there are different *software architectures*, requirements on the coordination of software components which need to interact using well-defined protocols to achieve the overall functionality of the system. A common example is a client-server architecture, where user clients mediate the user interaction and send requests to services on a server, which performs processing and responds with the results to the user. In order to recreate the functionality of the entire system, the reconfiguration of a number of interacting components into a common architecture will need to be recreated.
- **Diversity in scale of software.** Software ranges from individual routines and small programs which may only be a few lines long, such as Perl routines written for specific data extraction tasks; through products which provide particular sets of library functions; major application products, such as Microsoft Word, which provides a large group of related functionality to the user with large range of extra features, user interface support and backward compatibility; to large multi-function systems which provide entire environments or platforms for complex applications, such as the Linux operating system, which have millions of lines of code and entire sub-areas

which would be major products in their own right, but are required to work together as a coherent whole.

- **Diversity in provenance.** Software is developed by a wide range of different people organised in different ways. These would range from individuals writing specialised programs for personal use or to support particular functionality required by that individual; through community developments, where code is passed from person to person with an interest in developing further functionality; formal collaborative working as is widely undertaken in major open-source initiatives, such as Apache or Linux, where a mixture of diverse contribution to the core code base is combined with a more centrally controlled acceptance and integration procedure; to software developed and supported by a large or small team within a single organisation, for the internal purposes of the organisation, or else to be distributed usually as a commercial proposition. A single software product may pass through a number of different individuals and organisations with a number of different business goals, models, and licensing requirements. These different development models need to be reflected within attribution and licensing conditions.
- **Diversity in user interaction.** Software can support a wide range of interaction with the user. System software which controls the low level operation of the machine itself is designed to have no user interaction at all; library functions typically are designed to interact with other software components and have no or little user feedback, possibly delivering error messages; broader products are typically designed to have a user interface component which mediate commands from and responses to the user often via simple command-line or file based interaction. Other systems have rich user interactions with complex graphical user interfaces requiring keyboard and pointer and high-resolution displays, or audio input and output. Others require specialised input or output hardware devices such as joysticks and other control devices for games playing, or specialised screens and displays for virtual reality display. Clearly, in order to accurately reproduce the correct functionality of the software in the future, the appropriate level of user interaction will need to be recreated in some form.

Clearly there is huge diversity in the nature and application of software. However, we believe that there is sufficient commonality between these different scales that general principles for software preservation can be defined which are applicable to a wide range of different software products.

Further, although the word "preservation" implies availability for future generations, for software this can mean a much shorter time frame – much software has a working life (without additional modifications/rewriting) of five years at best and so curation can be considered to be quite a short term undertaking in the software domain. It is quite possible for the underlying data to have been created much earlier in time than the software used to manipulate it. It is the art of maintaining the underlying data by the way software systems are curated and migrated which is important in this sphere. There is a large overlap between preservation for the future and good practice for software maintenance in the here and now.

Thus there are a number of barriers and problems associated with preserving software. Consequently, although there are nevertheless good reasons to preserve software, there have been only limited consideration of the preservation of software as a digital object in its own right (see for example Zabolitzky:2002 for early work).

The work in this paper arose from a UK JISC sponsored study into the significant properties of software for preservation[1] (Matthews et al 2008), and subsequently in a JISC project into methods and tools for software preservation[2] (Matthews, Bicarregui et. al 2009). Given the relative immaturity of the field, the project developed a framework to express the notion of software preservation and set out some baseline concepts of what it means to preserve software. The framework has been developed further and deeper, bringing in notions from the OAIS reference model (OAIS 2002), and also developed tool support. The results of these two projects are reported in (Matthews, Shaon, Bicarregui, Jones, Woodcock & Conway 2009; Matthews, Shaon, Bicarregui, Jones & Woodcock 2009; Matthews 2010]. The benefits of software preservation were considered in a further JISC Study[3] between Curtis-Cartwright Consulting and the Software Sustainability Institute (Chue Hong et. al. 2010) (see chapter by Neil Chue Hong).

Software Preservation Approaches

Various approaches to digital preservation have been proposed and implemented, usually as applied to data and documents. However, they do usually refer to the means of preserving the underlying software used to process or render the data or document. Thus these preservation approaches directly relate to the preservation of software. The Cedars Guide to Digital Preservation Strategies (2002) defines three main strategies, which we give here, and consider how they are applicable to software.

- **Technical Preservation. (techno-centric**). This approach maintains the original software (typically a binary), and sometimes hardware, of the original operating environment. Thus this is similar to the use case for software preservation arising from museums and archives where the original computing hardware is also preserved and as much of the original environment is maintained as is possible. This is also an approach that is taken in many legacy situations; otherwise obsolete hardware is maintained to keep vital software in operation.
  Technical (hardware) preservation has the minimal level of intervention and minimal deviation from the original properties of the software. However, in the long-term this approach becomes difficult to sustain as the expertise and spare components for the hardware become harder to obtain.
- **Emulation (data-centric).** This approach recreates the original operating environment by programming future platforms and operating systems to emulate the original operating environment, so that software can be preserved in binary and run "as is". This is a common approach, undertaken in for example the PLANETS project[4], and the KEEP project[5], and also by groups such as the Software Preservation Society.
  The emulation approach for preserving application software is widespread, and is particularly suited to those situations where the properties of the original software are required to be preserved as exactly as possible. For example, in document rendering

[1] Joint Information Systems Committee (JISC) study into the *Significant Properties of Software* (2007). http://sigsoft.dcc.rl.ac.uk/twiki/bin/view

[2] Joint Information Systems Committee (JISC) sponsored project *Tools and Guidelines for Preserving and Accessing Software Research Outputs* (2007-09). http://www.stfc.ac.uk/e-Science/projects/medium-term/software-preservation/22426.aspx

[3] Joint Information Systems Committee (JISC) study ***Clarifying the Purpose and Benefits of Preserving Software. (2010-11)*** *http://softwarepreservation.jiscinvolve.org/wp/*

[4] http://www.planets-project.eu/

[5] http://www.keep-project.eu/

where the exact pagination and fonts are required to reproduce the original appearance of the document; or in games software where the graphics, user controls and performance (e.g. it should not perform too quickly for a human player on more up to date hardware) are required to be replicated. Emulation is also an important approach when the source code is not available, either having been lost or not available through licensing or commercial restriction. However, a problem of emulation is that it transfers the problem to the (hopefully lesser) one of preserving the emulator. As the platform the emulator is designed for becomes obsolete, the emulator has to be rebuilt or emulated on another emulator. Thus a potentially growing stack of emulation software is required[1].

- **Migration (process-centric).** Transferring digital information to new platforms before the earlier one becomes obsolete. As applied to software, this means recompiling and reconfiguring the software source code to generate new binaries, apply to a new software environment, with updated operating system languages, libraries etc.

  Software migration is a continuum. The minimal change scenario is that the source code is recompiled and rebuilt unchanged from the original source. However in practice, the configuration scripts, or the code itself may require updating to accommodate differences in build systems, system libraries, or programming language (compiler) version. An extreme version of migration may involve rewriting the original code from the specification, possibly in a different language. However, there is not necessarily an exact correlation between the extent of the change and the accuracy of the preservation.

  Software migration (or "porting" or "adaptive maintenance") is in practice how software which is supported over a long period of time is preserved. Long term projects such as Linux, or software houses such as Microsoft spend much of their effort maintaining (or improving) the functionality of their system in the face of environment change.

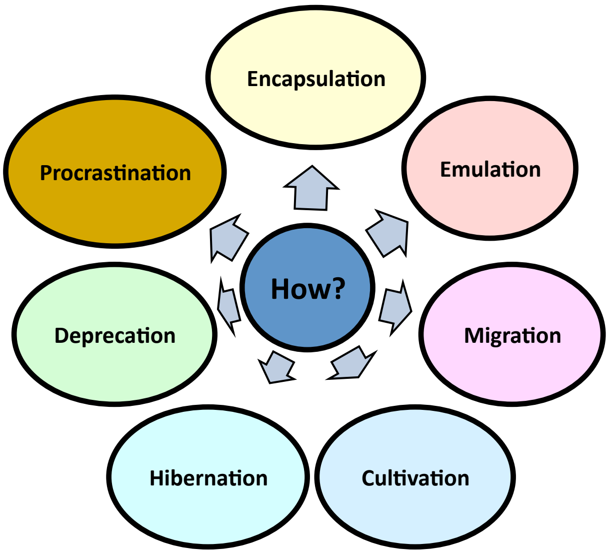


Figure 1. Seven approaches to Software Preservation.

[1] A problem addressed by the KEEP Virtual Machine, the bottom layer of which is simple enough to accommodate future emulators.

However, the migration approach does not seek to preserve all the properties of the original, or at least not exactly, but as observed in the European project CASPAR[1], only those up to the interface definition, which we could perhaps generalise to those properties which have been identified as being of significance for the preservation task in hand. Migration then can take the original source and adapt to the best performance and capabilities of the modern environment, while still preserving the significant functionality required. This is thus perhaps the most suited where the exact (in some respects) characteristics of the original are not required – there may be, for example, differences in user interaction or processing performance, or even software architecture – but core functionality is maintained. For example, for most scientific software, the accurate processing of the original data is key, but there is a tolerance to change of other characteristics.

These three approaches have been refined within the later JISC project to the seven different approaches given in Fig 1[2], which considered the social approach of the developers engaging in the preservation as much as the technical approach, for example distinguishing between migration, a notion of an "official" software port to provide a version on a new platform, and "cultivation", which involves managing the software in the long term by opening it up to a wider developer community, which can maintain and migrate the software.

In the rest of this paper, we are neutral to the preservation approach, but consider how the preservation of the key properties can be identified and checked.

Software Preservation Steps

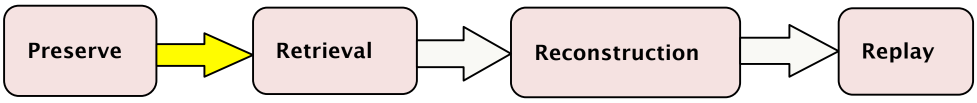


Figure 2. Steps in the Software Preservation process

Software preservation has four major steps, as in Figure 2.

- **Storage**. A copy of a software "product" needs to be stored for long term preservation. Software is a complex digital object, with potentially a large number of components constituting a product (an *information package* in OAIS); what is preserved is dependent on the software preservation approach taken. Whatever the exact items stored, there should be a strategy to ensure that the storage is secure and maintains its authenticity (*fixity* in OAIS terminology) over time, with appropriate strategies for storage replication, media refresh, format migration etc as necessary.
- **Retrieval**. In order for a preserved software product to be retrieved at a date in the future, it needs to be clearly labelled and identified (reference information in OAIS terminology), with a suitable catalogue. This should provide a search on its function (e.g. terms from controlled vocabulary or functional description) and origin (provenance information).
- **Reconstruction**. The preserved product can be reinstalled or rebuilt within a sufficiently close environment to the original that it will execute satisfactorily. For software, this is a particularly complex operation, as there are a large number of

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

[2] These are outlined at http://software.ac.uk/resources/approaches-software-sustainability Note that the authors refer to the approach labelled in this diagram as "Encapsulation" as "(technical) preservation".

contextual dependencies to the software execution environment that are required to be satisfied before the software will execute at all.

- **Replay**. In order to be useful at a later date, software needs be replayed, or executed and perform in a manner that is sufficiently close in its behaviour to the original. As with reconstruction, there may be environmental factors that may influence whether the software delivers a satisfactory level of performance.

In the first two steps, software (once a decision has been taken on what software components to preserve) is much like any other digital object type. Storage media that are secure and maintain integrity, and methods to identify and retrieve suitable objects are required in all cases. However, the problem of reconstruction and replay is especially acute for software. Digital objects designed for human inspection have rendering requirements that have issues of satisfactory performance; science data objects also typically require information on formats and analysis tools to be "replayed" appropriately. However, software requires an additional notion of a software environment with dependencies to other hardware, software, and build and configuration information.

Note that other digital objects require software to provide the appropriate level of satisfactory replay, and thus for other digital objects there is a need to preserve software too; as we shall see, there is also a dependency on the preservation of other object types (e.g. documentation) for the adequate preservation of software.

Thus in order to satisfactorily preserve software, we need to ensure that enough information is stored within the archival information package to support each of these steps in the future. We consider what kinds of information are required at each stage – see (Matthews, Bicarregui et. al 2009) for full details.

What do we need to support retrieval?

In order to support retrieval from a software library or archive, we need to have an indication of software in general terms to determine whether it is the right software component to meet the requirements of our upcoming task in the future. Thus we are likely to want to know the following items.

- **Gross functionality.** This would include a description of the purpose of the software product in general terms, with a description of its major inputs and outputs, together with a discussion of how the software operates. This may also include an overview of the general software architecture principles which the software operates (e.g. client-server architecture), and major dependencies.
- **Categorisation.** The software archive may support a categorisation under keywords or controlled vocabulary to support systematic searching of software.
- **Licensing.** There should be some indication of the ownership and legal control, and the licensing conditions under which the software is made available.
- **Provenance.** An indication of who developed the software for accurate attribution and possibly contact details for on-going support.

This information is that typically found on the "about" page of a software products website, and gives enough information to be able to discover the software in a search, determine whether it is likely to be of interest, and whether the user is entitled to use the software and under what conditions. Note that this information is independent of the software preservation approach used. Note also that this information does not include operating system information etc., which would typically be of interest to a user at this stage; strictly that would be reconstruction information below.

What do we Need to Support Reconstruction?

Software products typically come in many versions, which typically support different sets of functionality, and variants for different platforms[1]. Versions are associated with a release with specific functionality, and would often provide access to source code modules within specific programming languages, which would be provided with a build and install instructions to establish the version on a specific machine. A variant is associated with an adaptation of a version for a specific target environment. Usually it would be associated with an executable binary, but also could provide addition source modules that are tailored to the target environment. Thus in order to reconstruct the software, knowledge is needed of the precise environmental context in which the reconstructed software is expected to operate, so that dependencies can be tracked and satisfied on reconstruction. The exact information required to reconstruct the software would depend on the preservation approach (for example whether a binary executable is being emulated or a version rebuilt from source code, but we would expect that the following information would be needed:

- Set of component source code and/or binary files and their dependencies, including installation, configuration and build scripts and instructions as necessary. This would need to be the right files to provide the version required.
- Details of specific operating systems, versions and modifications.
- Details of programming language and versions, with appropriate versions of compilers to reconstruct the executable.
- Details of dependencies on other software, including auxiliary software libraries used to build the executable, and other software packages that are used in conjunction with software (for example, many programs use a database server supplied independently; the appropriate version and configuration of the database and possibly driver software would also need to be recorded).
- Details of hardware dependencies, including memory and processor requirements, screen resolution, and dependency on any special hardware peripherals.

Such material is what is usually supplied in the installation instructions supplied with software, and in well managed software development, the versions and their relationships would be maintained under a version management systems, such as CVS or Subversion.

As a consequence of this high dependency on the environment, preserving software is rarely a matter of preserving a single package of objects, but rather managing a collection of objects and their dependencies that need to be preserved in their own right.

What do we Need to Support Replay?

Once the software product has been satisfactorily reconstructed, then we need to know how to operate it. In this case, we need operating instructions and information on the expected inputs and outputs of the software. So the information needed to support replay would include:

- A detailed functional description of the operation of the software describing the actions it can undertake.

[1] Note that this use of the terms version and variant is ours.

- A description of the valid input formats, output formats, and how it handles error conditions.
- Details of the application programming interfaces it would support to interact with other programs.
- A description of the user interactions, in the form of manuals and tutorial materials and other usage documentation
- Any non-functional behaviour which the software is expected to support, such as response speed, data size, security

With this information, which is largely documentary, we can operate the software in the future. Note that documentary information is subject to the same preservation requirements as other documents, requiring the appropriate rendering software and the preservation of properties significant to that document format.

How do I Judge now that what I have Preserved is "Enough"?

We have discussed that software preservation requires that the three stages of retrieval, reconstruction and replay must all be satisfactorily supported in order to provide a future user with the capability to use the software. But how do we know when we have provided enough information so that the resulting software product performs "correctly", that is in the manner the original developers intended?

We introduce a notion of *performance* to demonstrate that a particular reconstruction adequately preserves the required characteristics of software. Performance as a model for the preservation of digital objects was defined by the National Archives of Australia in (Heslop et. al. 2002) to measure the effectiveness of a digital preservation strategy. Noting that for digital content, technology (e.g. media, hardware, software) has to be applied to data to render it intelligible to a user, they define a model as in Fig 3 *Source data* has a *Process* applied to it, in the case of digital data some application of hardware and software, to generate a *Performance*, where meaning is extracted by a user. Different processes applied to a source may produce different performances, and it is the properties of the performance that need to be considered to assess the value of a preservation action. And the properties of the performance can arise from a combination of the properties of the source with the technology applied in the processing.

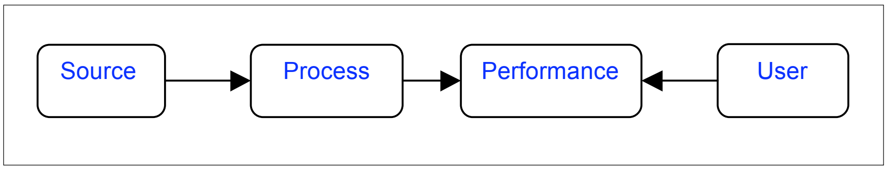


Figure 3. The NAA Performance Model

The notion of performance has been developed in the context of traditional archival records to identify the significant properties of different media types which compare the performance created by the original process of rendering with that created by later rendering processes on new hardware and software. The question that arises is how this model applies to software itself.

In the case of software, the performance is the execution of binary files on some hardware platform configured in some architecture to provide the end experience for the user. However, the processing stage depends on the nature of the software artefacts preserved which have differing reconstruction and replay requirements.

- In the case where binary is preserved, the process of generating the performance is one of preserving the original operating software environment and possibly the hardware too, or else emulating that software environment on a new platform. In this case, the emphasis is usually on performing as closely as possible to the original system.
- When source code and configuration and build scripts are preserved, then a rebuild process can be undertaken, using later compilers and linkers on new a new platform, with new versions of libraries and operating systems. In this case, we would expect that the performance would not necessarily preserve all the properties of the original (e.g. systems performance, or exact look and feel of the user interface), but have some deviations from the original.
- In an extreme case, only the specification of the software may be preserved. In this case, a performance could be replicated by recoding the original specification. Here, we would expect significant deviation from the original and perhaps only core functionality to be preserved. This case would seem to be exceptional. However, it is less unusual in coding practice, as products are often migrated into a different language; for example the NAG library originated in FORTRAN, but later produced a C version. In some circumstances, this is a result of *reverse engineering* where source code (or even in extreme cases binary code) is analysed to determine its function and recoded.

A software performance can thus result in some properties being preserved, and others deviating from the original or even being disregarded altogether. Thus in order to determine the value of a particular performance, we define the notion of ***Adequacy***.

> *A software product (or indeed any digital object) can be said to perform adequately relative to a particular set of features ("significant properties"), if in a particular performance (that is after it has been subjected to a particular process) it preserves that set of significant properties to an acceptable tolerance.*

This notion of adequacy is usually viewed as an aspect of the established notion of ***Authenticity*** of preservation (i.e. that the digital object can be identified and assured to be the object as originally archived). However, we feel that it is useful to separate these two notions in order to establish a more lucid requirement specification of long-term preservation of software. For this, we use the premise that the term ***Authenticity*** in long-term preservation essentially signifies the level of **trust** between a preserved software product and its future end users. From the perspective of an end user of a software product, this trust is primarily associated with the ability to trace the provenance (e.g. history of origin, custodianship etc.) and verify the fixity information (e.g. checksum) of the software. For example, a preserved software product with comprehensively documented provenance history and verifiable fixity information might establish a sense of trust for the body responsible for its preservation in its users. But this "trusted preservation" does not guarantee a reliable behaviour from the software once reconstructed in future; it might incur a loss of some of its original features during its reconstruction process. However, the software could still be used for the remaining features retained after reconstruction, which could be sufficient to extract an acceptable level of performance from the software. An example of such software is the emulated version of the 1990's DOS-based computer game Prince of Persia[1]. While some of the instructions do not always work on the emulator and the original appearance of the game is also somewhat lost, it is possible to run the emulator to play the complete game on a contemporary computer platform. The term ***Adequacy*** introduced here is intended to represent this particular concept. In effect, by measuring the adequacy of the performance, we can thus determine how well the software has been preserved and replayed.

The distinguishing feature of the performance model for software is that the measure of adequacy of the software is closely related to the performance of its input *data*. The purpose of software is (usually) to process data, so the *performance* of a software product becomes the *processing* of its input data. This relationship is illustrated in Figure 4. Note that we have reversed the arrow between performance and user to reflect the information flow. Further, there is an interaction between the user and the software performance, reflecting the user's interaction with the software product during execution, changing the data processing and thus the data performance.

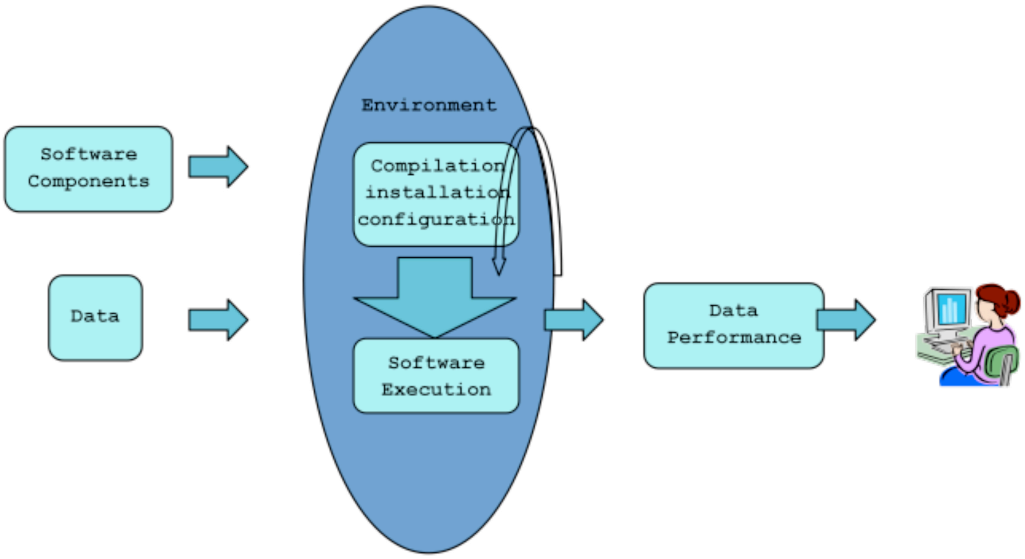


Figure 4. Performance model of software and its input data.

For example, in the case of a word processing product that is preserved in a binary format, which is processed via operating system emulation, the performance of the product is the processing and rendering of word processing file format data into a performance which a (human) user can experience via reading it off a display. The user can then interact with the processing (via for example entering, reformatting or deleting text) to change the data performance. Thus the measure of adequacy of the software is the measure of the adequacy of the performance when it is used to process input data, and also preserving a known change in the data performance that results from user interaction with the processing.

The adequacy of different preservation approaches is dependent upon the performance of the final result on the end use on *data*. As the software has to be able to produce an adequate performance for any valid input date, the adequacy can be established by performing trial executions against representative test data covering the range of required behaviour (including error conditions) against the significant performance property. As a

[1] http://www.bestoldgames.net/eng/old-games/prince-of-persia.php

consequence, it is also necessary to preserve test data to establish adequacy as part of the software preservation process.

This notion of performance can be applied recursively to software that processes other software, for example software used for emulation or compilers to build software binaries, which also need to be preserved, as in Figure 5. In this case, the performance of software is the processing of the application binaries or source code, which in turn is measured by its adequacy in processing its intended input data. Figure 5. also illustrates that the stages in the performance model can be related to the stages in the software preservation process, and the information at each stage to data in the OAIS information model. For more details see (Matthews, Bicarregui et. al 2009).

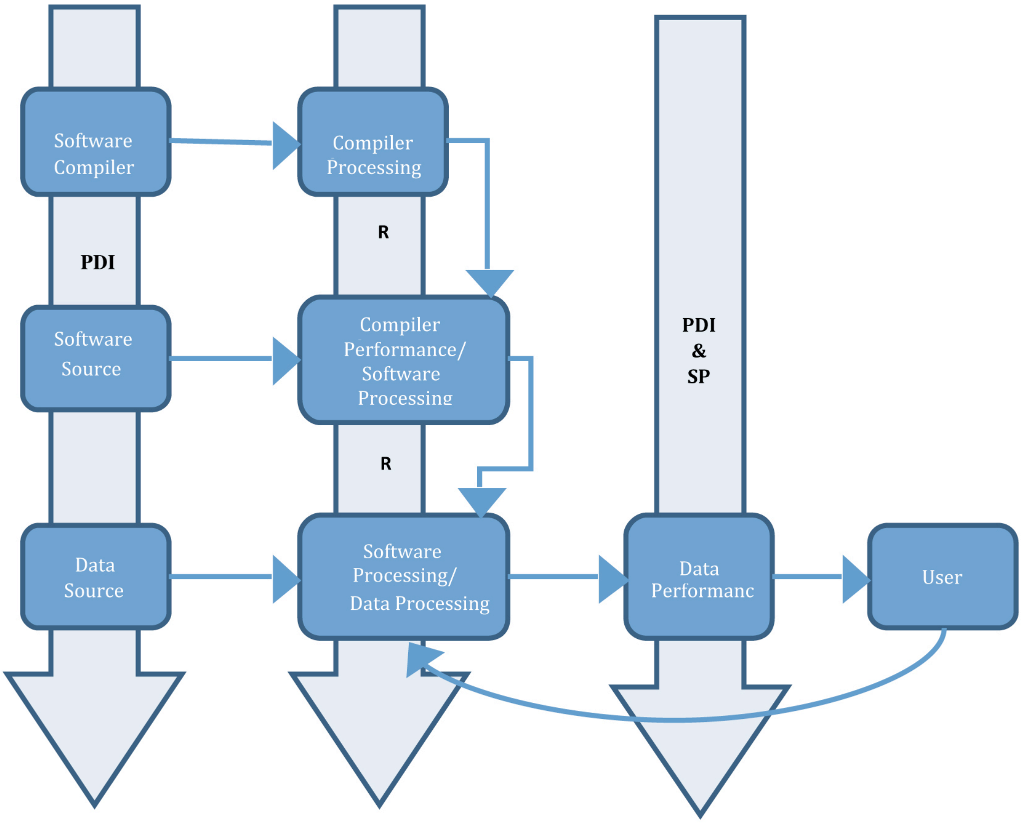


Figure 5. Applying the software performance model to software that processes software.

Some Examples of Tests of Adequacy

To illustrate how adequacy of software may be tested, below we give an indication of the likely tests which would be needed to be maintained for different categories of data, outlining some probable adequacy determining factors for a number of different categories of software products:

- **Scientific Software.** Scientific analysis software is usually most concerned with maintaining the accuracy of the calculations. Thus such adequacy conditions as *the system should calculate the Fast Fourier Transform*, or *the result must be accurate to 8 decimal places* would be appropriate. Other factors such as user interfaces and control may be less crucial to be preserved and changes may be tolerated while maintaining adequacy. So typically, for scientific software, adequacy is established by running the software to process some pre-specified test input data, comparing the output of the test run with the corresponding pre-specified test result, and checking if the output exceeds the acceptable level of error tolerance for the software.

For example, the NAG library[1] publishes test cases and a specification of the required accuracy, in terms of number of significant figures for its mathematical software routines. Such accuracy is highly dependent on the mathematical libraries and coprocessor used in reconstruction and replay.

- **Games Software.** There is considerable interest particularly within the hobbyist community to maintain games from obsolete hardware[2]. In this case, the emphasis is on playability, so that the user graphics and user controls are the key factors. Note in this case, we may need to "dumb-down" highly performant modern hardware to the more limited capability of the past. So in this case, adequacy is established by comparing its User Interface (UI) with the screen capture of its original UI, possibly including video, and comparing its performance against some pre-defined use cases. For example, the completion time of a particular level can be compared against the average completion time for that level in the original game.
- **Programming Language Compilers.** In the case of a compiler, we would need to check that it covers all features of the programming language that it supports, e.g. concurrency (i.e. threads), polymorphism, etc. For some programming languages (e.g. Fortran, C, C++ etc.), there exist ISO standards[3] that describe the correct behaviour of a piece of software written in these languages. These standards also provide test suites that may be used to assess the adequacy of a compiler for rendering all features of the programming language that it supports. As noted above, this would need to ensure that the application resulting from compiling its source code (written in a language supported by the compiler) using the compiler yields the expected behaviour.
- **Word Processing software.** The adequacy of a word processor may be measured based on its ability to render existing supported word documents with an acceptable level of error tolerance. For example, a word processor may be regarded as adequate as long as it clearly displays the contents (e.g. test, diagram, etc.) of a word document, even if some of the features of the document content, such as font colour and size, may have been rendered incorrectly or even lost completely. However, in other circumstance, fonts, colours and page layout may be considered a key property to be preserved. Furthermore, the processing software would need to enable editing (e.g. add/change/remove text, change font) and saving existing word documents, and enable creation and saving of new documents.
- **Digital artworks.** In the case of digital artworks, there are two judges of adequacy; the intention of the artist and the experience of the audience. The software may change and evolve over time as it is migrated to new platforms in the curatorial process, and it is arguable that as long as the user experience is measurably the same, the artwork can be said to be preserved. However, the intent of the artist is also important. Artists, who often may be the programmers of the work, may consider the code as a core part of their artistic input. Thus they may consider the form of the code as a significant property in

[1] http://www.nag.co.uk

[2] See http://www.softpres.org/, who have adopted the software preservation term specifically for this purpose.

[3] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_tc_browse.htm?commid=45202

its own right, which should be adequately preserved and resist attempts to change it as part of a migration strategy.

Supporting Software Preservation in Practice

In this paper, we have presented some of the issues and concepts that are involved in the preservation of software. However, applying these in practice remains a complex and challenging undertaking. In this section, we consider one case study and some tools and methodologies that can help mitigate these challenges.

The BADC case study

The National Centre for Atmospheric Science's British Atmospheric Data Centre (BADC) is a NERC Designated Data Centre that currently has over 250TB of atmospheric data and are currently archiving some 200 datasets for the consumption of UK scientists and researchers[1]. In order to facilitate efficient accessibility and usability of these large volumes of atmospheric data, the BADC also provides access to a variety of software, which ranges from very simple data conversion tools to highly complex weather prediction models. Considering the importance of the BADC software in enabling accessibility and interpretation of its large data holdings, effective long-term preservation (i.e. reuse) of the BADC datasets implies the need for appropriate preservation actions for its software.

We consulted with the BADC software development and maintenance team to analyse their approaches to software maintenance in the context of a number of their software products, such as the BADC trajectory service. This indicated that long-term preservation of these software products is not considered within the current operational remit of the BADC. On the whole, the BADC considers the long term archiving of software an impractical option principally due to the complex dependencies of software. It takes the view that it expects much current software will be superseded by newer software that will be capable of recreating and enhancing much of the existing analysis and access functionality. Furthermore, the BADC considers the costs of preserving software, especially in terms of migrating to newer technological platform, to be a prohibitive factor and hence outside their current remit.

In general, the BADC case study reinforces the view that the inherent complexity of software is a significant barrier to its long-term preservation. In addition, it highlights another prohibitive factor for long-term preservation of software: the *cost of preservation*. Effective preservation of a digital object over the long-term requires its continuous management and enhancement over its lifecycle. This involves, amongst other tasks, periodically assessing (and improving) the preservation strategy to ensure the effective re-construction and re-use of the digital object notwithstanding any related technological changes. This is expected to impose significant recurring costs on the organisation undertaking long-term preservation, in terms of technical resources, personnel effort etc. required. There is likely to be even greater effort and hence costs required for the long-term preservation of software due to the complex dependencies between its components. Thus, for an organization which is already bearing the costs of maintaining and

[1] http://badc.nerc.ac.uk/home/index.html

developing a wide array of software, it would be difficult to justify and incorporate within its current remit and budget, the additional costs of long-term software preservation as such an activity might not be deemed beneficial to BADC in the short-term. In addition, the high complexity and costs of employing currently available preservation mechanisms, such as migration and emulation would also add to the overall costs of long-term preservation of software.

Therefore, we envision that the organisations, such as the BADC should benefit from the conceptual framework for software preservation presented in his paper. The framework provides a comprehensive and organised view of the underlying dependencies of software in the context of preservation and facilitates accurate identification of the software properties needed for its effective preservation. This can aid in efficient management of the complexity of software preservation, which in turn could help reduce the overall costs of preservation. Additionally, the framework could potentially be used for incorporating long-term preservation functions into existing systems for software development and maintenance, such as the one at BADC.

Evaluating the Software Preservation Framework against BADC Software

We have evaluated the framework against a number of pieces of BADC software, such as the BADC Web Feature Service (WFS)[1]. For this, we tried to collect the appropriate value(s) for each of the preservation properties defined in the framework for each major conceptual entity of software.

The experience of applying the framework for software preservation to the BADC software has shown that the framework is sufficiently relevant to the software used as well as being adequate in terms of the information recorded. However, it has also highlighted the necessity to have considerable knowledge of both the framework and software in question to accurately apply the framework to the software. This indicates a need for tools to facilitate the recording of software preservation properties by providing guidelines that, for example, explain the underlying concepts of the framework in a user-friendly manner.

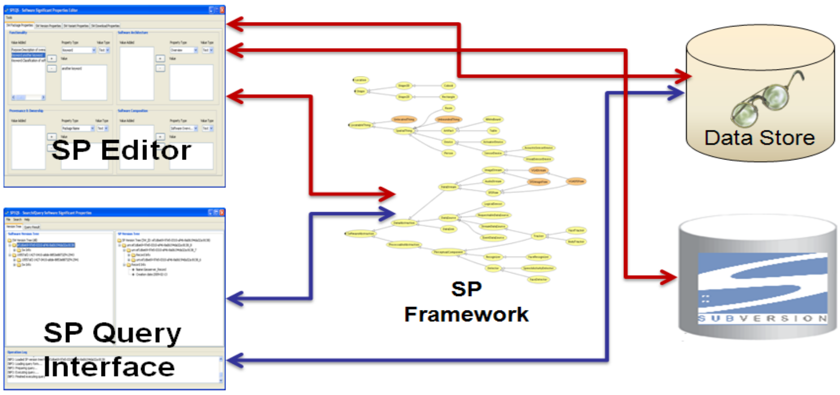


Figure. 6. The SPEQS software architecture.

[1] The BADC WFS Enables retrieving and updating geospatial data encoded in Geographic Markup Language (GML – http://www.opengeospatial.org/standards/gml), or any GML-based formats, irrespective of the location or storage media of the data. The implementation is based on the Open Geospatial Community (OGC) standard for Web Feature Service(http://www.opengeospatial.org/standards/wfs)

Consequently, we have developed a tool, namely **Significant Properties Editing and Querying for Software** (SPEQS), illustrated in Figure 6. that demonstrates the feasibility of incorporating capturing preservation properties of software within the software development lifecycle to aid its long-term preservation. It has been implemented in Java as a plug-in for Eclipse, a widely used Open Source interactive software development environment, to enable software developers to record, edit and query preservation properties of software directly from within the Eclipse environment. This approach of enabling the developer(s) of a software project to record its preservation properties is envisaged to contribute towards ensuring the accuracy of the information recorded.

Software in Preservation Planning

Clearly, with such a complex network of dependencies and the need to establish adequacy conditions in advance, software artefacts need to be considered carefully in the cost-benefit analysis and planning of a preservation objective. As a consequence, we have been considering the role of software in Preservation Network Models (PNMs). PNMs were originally developed within the CASPAR project and are described in detail in (Conway, Dunckley et al., 2009; Conway, Giaretta et al, 2009) as a formal model for conceptualising the relationships between resources within the scenario of a preservation objective, and can be used for planning and monitoring preservation scenarios for particular resources.

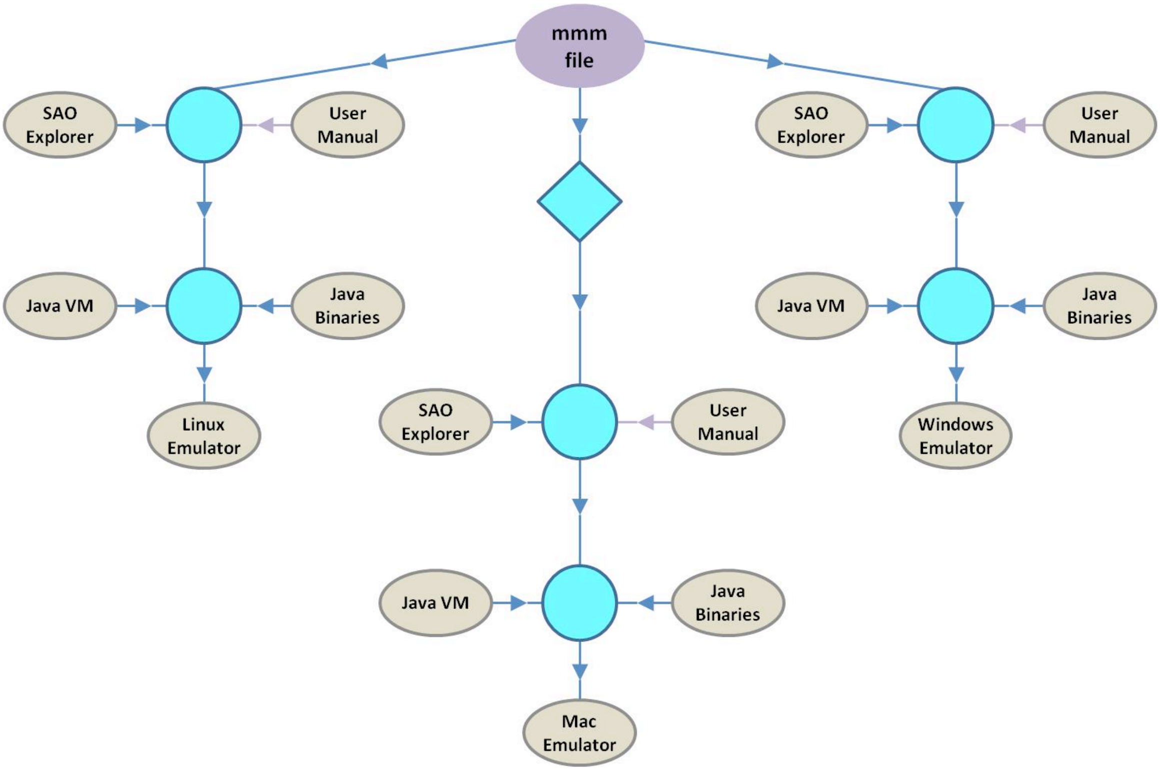


Figure 7. A PNM presenting the preservation of software with alternative strategies

Figure 7 presents a preservation network model for data that represents an ionization profile of the atmosphere. The network shows alternate strategies for preserving a mmm "ionsonde"[1] file for platform based variants. In this case, we show three alternate strategies for preserving the SAO-Explorer[2] software to interpret the data, the choice of branch depending on the underlying platform (Linux, MacOs, Windows) available to the user. Each branch presents the items requiring preservation, consisting of the SAO-Explorer executable file for the platform and its user documentation, and an extension through the software stack to record the dependencies of the software. In this case, we need to preserve the Java VM and software platform to maintain the executability of the SAO-explorer within each respective operating platform. By doing so we preserve the ability to process the data in the file, apply standard analysis model to and then render it in the form of an ionogram.

The use of such preservation networks to plan and monitor preservation strategies, including that of software, and to manage preservation risks, is the subject of current research within the SCAPE and ENSURE projects (Conway et al, 2011).

Conclusion

In this paper we have presented some of the issues around preserving software and put forward a conceptual framework to express a rigorous approach to long-term software preservation. We believe that this is a general and principled approach which can cover the preservation needs of a wide range of different software products (e.g. the BADC software products), including modern distributed systems and service oriented architectures, which are typically built of pre-existing frameworks and have a large number of dependencies on a widely distributed network of services, many of which are outside the control of the typical user (e.g. DNS services, proxies). We also believe that the performance model presented here, which introduces a notion of ***Adequacy*** of software performance as well as a notion of user feedback to influence the performance, represents an approach to preserving the user interface and the user interaction model, although work is required to further develop that notion.

Current work is underway to incorporate the preservation of software within preservation planning and managements, so that it can be managed in a cost effective manner with the preservation of other digital artefacts. In addition, further work is required to evaluate the preservation framework, especially against a range of software types to cover the diversity of software and to consider how to support the preservation of legacy software.

As a general principle, software preservation, should ideally be considered within a software engineering process. Software developers, driven rightly by the immediate needs of developing and maintaining software, rarely consider the implications of the long term preservation of their software. However, many of the disciplines which are part of good software development practice also make the task of preserving software much more tractable. These would include: good version control; good configuration and build scripts and processes; good documentation for both the developer and user; and systematic and well documented testing and test cases which both ensure program correctness and are also the key to assuring adequacy of preservation. Further, many of the tools and techniques which promote software reuse in software engineering are also

[1] An Ionosonde is an instrument that examines the ionosphere by means of high frequency radar; the resulting data is often presented in an isonogram that graphs the virtual height of the ionosphere against frequency.

[2] SAO Explorer http://ulcar.uml.edu/SAO-X/SAO-X.html

applicable to software preservation Thus, we could say that good software engineering leads to good software preservation

Acknowledgments. The work in this paper was carried out under the JISC study into the Significant Properties of Software and also the JISC sponsored Tools and Guidelines for Preserving and Accessing Software Research Outputs. Further discussions were undertaken with Curtis+Cartwright and the Software Sustainability institute in the JISC Benefits study. in the Further work has continued within the CASPAR, ENSURE and SCAPE European projects. We would like to thank our colleagues Juan Bicarregui, Catherine Jones, Jim Woodcock, David Giaretta, Steven Rankin, Matt Dunkley, Simon Lambert and Michael Wilson for their advice and discussions.

References

The Cedars project. The Cedars Guide to Digital Preservation Strategies (2002). 16 pp. Retrieved from http://www.imaginar.org/dppd/DPPD/146%20pp%20Digital%20Preservation%20Strategies.pdf

Chue Hong, N., Crouch, S., Hettrick, S., Parkinson, T., & Shreeve, M. (2010). Software Preservation Benefits Framework. 74 pp. Retrieved from http://www.software.ac.uk/attach/SoftwarePreservationBenefitsFramework.pdf

Conway, E., Dunckley, M., Giaretta, D., & McIlwrath, B. (2009). *Preservation Network Models: Creating Stable Networks of Information to Ensure the Long Term use of Scientific Data*. Paper presented at the Proc. Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data (PV 2009), Villafranca del Castillo, Madrid, Spain. http://www.sciops.esa.int/SYS/CONFERENCE/include/pv2009/papers/6_Conway_PreservationNWModelling.pdf

Conway, E., Giaretta, D., Lambert, S., Dunckley, M., & Matthews, B. (2011). Curating Scientific Research Data for the Long Term: a Preservation Analysis Method in Context. *International Journal of Digital Curation, 6*(2), 38-52. Retrieved from http://epubs.stfc.ac.uk/bitstream/7067/Conway-IJDC-204-861-4-PB.pdf

Conway, E., Matthews, B., Giaretta, D., Lambert, S., Draper, N., & Wilson, M. (2011). Managing Risks in the Preservation of Research Data with Preservation Networks. Paper presented at the 7th International Digital Curation Conference (IDCC2011), Bristol, UK, 05-07 Dec 2011. http://epubs.stfc.ac.uk/bitstream/7171/Final_ManagingRisks_Conway_IDCC11.pdf

Heslop, H., Davis, S., & Wilson, A. (2002). National Archives Green Paper: an Approach to the Preservation of Digital Records. Retrieved from http://nla.gov.au/nla.arc-49636

Matthews, B. M., Bicarregui, J. C., Shaon, A., & Jones, C. M. (2009). A Framework for the Significant Properties of Software. JISC Tools and Methods for Software Preservation Project report. Retrieved from http://epubs.stfc.ac.uk/work-details?w=51076

Matthews, B. M., McIlwrath, B., Giaretta, D., & Conway, E. (2008). The Significant Properties of Software: A Study, 99 pp. Retrieved from http://www.jisc.ac.uk/whatwedo/programmes/preservation/2008sigprops

Matthews, B. M., Shaon, A., Bicarregui, J. C., & Jones, C. M. (2010). A Framework for Software Preservation. *International Journal of Digital Curation, 5*(1), 91-105. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/148

Matthews, B. M., Shaon, A., Bicarregui, J. C., Jones, C. M., J., W., & Conway, E. (2009). *Towards a Methodology for Software Preservation*. Paper presented at the 6th International Conference on Preservation of Digital Objects (iPRES 2009) San Francisco, USA, 05-06 Oct 2009. http://epubs.stfc.ac.uk/bitstream/4599/IPres2009_Matthews_swpres.pdf

Matthews, B. M., Shaon, A., Bicarregui, J. C., Jones, C. M., & Woodcock, J. (2009). *An Approach to Software Preservation*. . Paper presented at the Proc. Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data (PV 2009), Villafranca del Castillo, Madrid, Spain, 01-03 Dec 2009. http://epubs.stfc.ac.uk/bitstream/4790/PV2009_37_Shaon_ApproachToSWPreservation.pdf

OAIS. Reference Model for an Open Archival Information System. Recommendation for Space Data Systems Standard, CCSDS Blue Book. (2002). Retrieved from http://public.ccsds.org/publications/archive/650x0b1.pdf

Zabolitzky, J. G. (2002). Preserving Software: Why and How. *Iterations: An Interdisciplinary Journal of Software History, 1*. Retrieved from http://www.cbi.umn.edu/iterations/zabolitzky.html

The Villa of Oplontis: A "Born Digital" Project

John R. Clarke

Department of Art and Art History, College of Fine Arts, University of Texas at Austin, 1 University Station #D1300,
Austin, TX 78712-0337 USA

Abstract. The Oplontis Project has as its goal the definitive study and publication of Villa A ('of Poppaea') at ancient Oplontis (Torre Annunziata, Italy), built in 50 B.C. and destroyed by the eruption of Vesuvius in A. D. 79. Two recent developments, the born-digital, XML E-Book, and the gaming engine Unity, have made it possible to surpass the limitations of the print monograph. The 3D model produced by King's Visualisation Lab is linked both to the E-Book and the Project Database; it acts as an index to the information gathered by the 42 scholars contributing to the four-volume publication. A viewer can click on features to enter the database, where high-resolution images accompanied by a fully scholarly apparatus reside. The 3D model allows reconstructions of lost features of the ancient Villa, including its geological setting, the placement of its sculpture and plant materials, destroyed upper stories, and orphaned fragments of wall painting.

Introduction

In 2005, the Centre for the Study of Ancient Italy at the University of Texas entered into a collaboration with the Archaeological Superintendency of Pompeii, a branch of the Italian Ministry of Culture, to study and publish one of the largest ancient Roman luxury villas buried by the eruption of Vesuvius on 24 August A.D. 79. Known officially as Villa A at Torre Annunziata—the modern town built on top of the ancient town of Oplontis, the Villa of Oplontis lies under 8.5 m. of volcanic material, and is about three miles north of Pompeii. Its importance rests on three facts: it is enormous, with 99 excavated spaces; its decoration is exquisite; and it may have belonged to Nero's unfortunate second wife, Poppaea Sabina.

The Work Done

To develop a research strategy, I assembled a small group of experts, including an architect, a photographer, an art historian, and an archaeologist. Our questions included: How did the current reconstruction of the Villa come about? What can archives and excavated artefacts tell us about the Villa? What lies beneath it? What are the meanings of its vast and complex decorative apparatus? How did its residents, including the masters, guests, and slaves, use the Villa? To answer these questions we established the Oplontis Project team, and embarked on a six-year campaign of research and excavation (Oplontis Project website 2011[1]).

Since the principal goal of the Oplontis Project is the definitive publication of Villa A, the question was how to publish. Over the past thirty years, two acclaimed print series had set the standard: the German series, *Häuser in Pompeji*, or *Houses in Pompeii,* and the four-volume publication of the Insula of the Menander at Pompeii edited by Roger

[1] http://www.oplontisproject.org/index.html

Ling. Both of these print series aims to document the houses in question to the fullest possible extent: architecture and construction techniques, decorations—including pavements and wall and ceiling paintings—statuary, small finds, and much more.

Presentation of such complex materials is a problem. Even in the folio format used by the *Häuser im Pompeji* volumes, illustrations often represent a large wall painting inadequately (Strocka 1991). To create the drawings that illustrate all the walls of the house, whether decorated or not, draftspersons actually traced the walls on huge sheets of mylar. They then redrew them, using various graphic devices to make them legible. Despite the great pains taken, these drawings remain schematic. To complete lacunose wall decorations, the graphic artists propose conservative reconstructions, drawn into the actual-state tracing in faint outlines. Black-and-white photographs document the actual state of each wall. Often, especially in narrow spaces, the actual-state photographs must be taken from an oblique angle, making them difficult to understand.

In the smaller-format volumes for the Insula of the Menander publication, the editor often utilizes long gatefolds to show—again in a schematic way—the positioning of the decorations of their respective walls (Ling 2005). In order to save space and paper, drawings from different spaces sometimes appear together on the same gatefold. A scholar wanting to investigate the decoration of a particular room must first find its number on the plan at the back of the book, then consult the catalogue description, then find the appropriate foldout, then go to a separate plates section to find the color illustration, and to yet another section to find the black-and-white photos. She will also have to take the scale markers seriously, for the drawings are reproduced at different sizes.

Was there a better way to publish Villa A at Oplontis? Since its inception in 1999, I had been following the successes of the Humanities E-Book series published by the American Council of Learned Societies in New York (2011). The ACLS, created in 1919, represents 81 societies in the humanities and the social sciences, providing leadership and funding for scholarly initiatives. The Humanities E-Book is a digital collection of over 3,000 full-text titles offered by the ACLS in collaboration with 20 learned societies, nearly 100 contributing publishers, and librarians at the University of Michigan's Scholarly Publishing Office. The result is an on-line, fully searchable collection of books in the Humanities that have been recommended and reviewed by scholars. The back-list collection, consisting of sought-after, out-of-print titles, consists of scanned books. The subscriptions to these back-list titles provide funding for the front-list titles, currently numbering 88. The ACLS commissions the front-list titles from established scholars who wish to take full advantage of the possibilities of digital publication. These are the so-called "born-digital" e-books. In contrast to books that are image-based, the 88 XML books are text-based, a format that allows for features such as image enlargement, internal and external links, interactive maps, audio and video files, databases and archival materials. All of these are things that the Oplontis Project publication wants to do.

The publication is complex. At this point, we have 42 authors whose contributions will cover, in four volumes: the ancient setting and modern rediscovery, the decorations, the excavations, and the architecture of the villa. The current plan shows the 99 spaces, including beautifully painted rooms, gardens, and a 61 metre swimming pool, including our excavations since 2006, consisting of twenty trenches in all (Figure 1). In addition, we need to document the excavations from 1783 through 2010 on the basis of texts, drawings, and photographs. We have 19 large-scale sculptures, marble columns and capitals, ceramics, and bronzes. And we have to make clear what is ancient and what is modern.

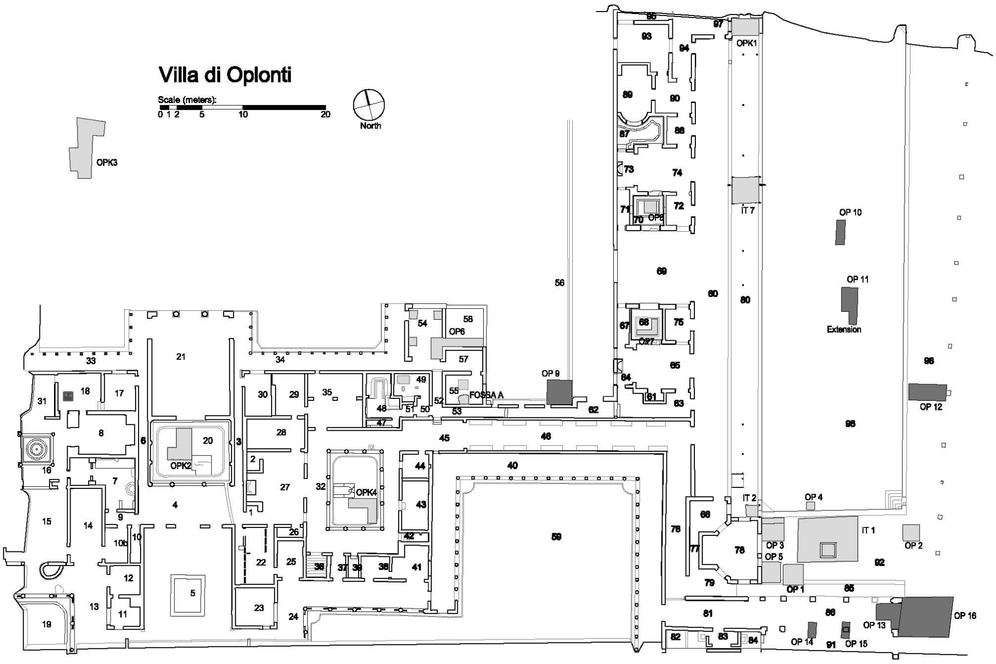
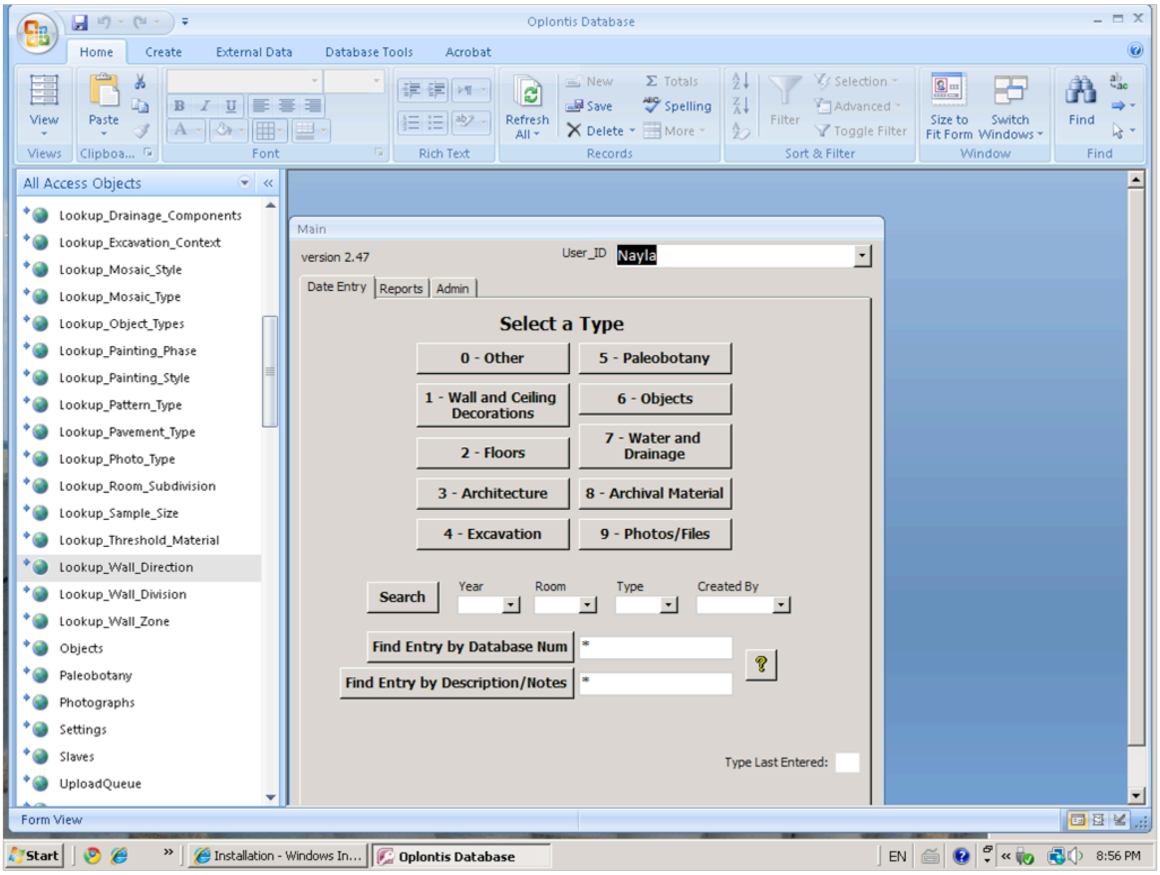


Figure 1. Villa A, Torre Annunziata, plan 2010. Jess Galloway

We have to organize this mass of information for a born-digital e-book. This is where the digital model comes in. Our goal is to take the finding tools of the print book and locate them within a navigable, 3D model of the Villa of Oplontis. Scholars agree that the effectiveness of the scholarly print book rests of the accuracy and effectiveness of the finding tools: the table of contents, the index, figure call-out numbers, notes, and—above all—cross references. All of these tools allow a reader to find information. From the author's point of view, these tools are the building blocks of his or her argument. They allow the author to take the reader from text to image and back, to compare one set of information (such as bibliographical references or visual comparisons) with another set of information. Clearly, recasting these traditional finding tools in the born-digital environment means faster access to a much greater quantity of information, all of it located in the actual spaces of Villa A. In short, the 3D model becomes the index to the digital book.

To make the model function as the index, we had to create a database, and make access to the database possible by clicking on images or spaces within our index—that is, within the digital model itself. We have spent considerable time and effort refining the database to make it represent our research well. The fields include the major types of information we have. The platform is Microsoft Access, and all of the fields are searchable by keyword (Figure 2). Most importantly, it allows us to upload high-resolution photos of each item, and it automatically associates photographs of items in categories 1 through 8 with category 9, photographic archives. The database assigns each item a unique number. For archaeological finds, that unique number becomes the inventory number; the same is true of any item that the Oplontis Project catalogues and describes: wall paintings, whether in situ or unassociated with a specific wall; archival photographs, plans, and drawings; documents such as the excavation journals. The full description appears in a searchable field; a thumbnail image appears at lower right. Clicking on the thumbnail allows the user to download a high-resolution image. An important feature of the database is that when there exist many images of the same artefact—say of a wall painting discovered 45 years ago—the database will associate all the images of that artefact contained in the database. This essay demonstrates how important photographic archives are to the success of this project.



version 2.47

User_ID Nayla

Figure 2. Screenshot of Oplontis Project database entry page

Although from the beginning we knew we wanted to publish a born-digital e-book, it was in the third year of our work that we met with Richard Beacham and decided to partner with the King's Visualisation Lab. Our collaboration found success in two substantial grants, one a collaborative research Fellowship from the National Endowment for the Humanities, and on the British side, a grant from the Leverhulme Foundation. As the reviewers for both of these granting agencies acknowledged, it was important, as part of the e-book publication, to create a 3D digital model. In addition to its role of disseminating the results of our studies of the ancient Villa of Oplontis, the model must also preserve the Villa itself for posterity. In light of the severe damage the Villa has suffered from exposure to the elements and improper care, a digital actual-state record is essential for all future work. Creating this model from the resources at hand has proven an exciting and often daunting task.

We began with the idea of scanning all 99 spaces of the villa using available laser-scanning technology, but this proved too costly. We then decided to build the model in 3D Studio Max, based on the electronic plans and sections created by our architects. We assigned one of our architects, Tim Liddell, to provide resources for KVL's Drew Baker, so that together we could make an accurate 3D Max model. The Oplontis Project engaged an experienced architectural photographer, Paul Bardagjy, to create the most accurate possible photographs of walls, floors, ceilings, and colonnades—and to oversee stitching the photographs to scale. The stitching proved particularly demanding and time-consuming. Many long spaces, such as the Villa's porticoes (33, 34, 24, 60) had to be stitched together from numerous shots, required the work of several professionals to create an accurate, undistorted record of just what is there. Because of the scholarly nature of the project, reproduction of the actual state of the Villa remained the fundamental demand of our publisher, and an absolute criterion for our professional readers: archaeologists, architects, and art historians.

With much work, determination, and considerable expense, we have arrived at the alpha-state model. Three features are especially exciting from the point of view of the digital humanities. First, the Unity interface allows real-time exploration of all the spaces of the Villa. Second, the user can click on features to find out more about them. Third, the user can switch from the actual-state model to accurate reconstructions of complex rooms. These three features, we believe, put new tools in the hands of scholars and the public alike.

In what follows, I wish to highlight one particular aspect of the e-book: the use of archival material. The most straightforward kind of inquiry that the model can answer is what the villa looked like in antiquity. Although they might not know it, the structure that visitors see when they come to the site at Torre Annunziata is largely a fabrication. When modern excavations began in 1964, the aim of the Italian government was to make the Villa of Oplontis into a tourist site. To make it look like a "real" ancient Villa, contracting companies were called in to use any means available to turn the rubble into a living museum. This photo of the unearthing of the north portico of the Villa (33) demonstrates the enormity of this task (Figure 3). We see the columns cut down by a pyroclastic surge, made of superheated volcanic mud that sped along the ground at speeds up to 650 miles an hour. These surges cut down everything in their path. The construction company used the remaining pieces to make whole columns and piers that to support new architraves and roofs. It is the duty of the Oplontis Project to provide photos like these to demonstrate the true nature of the artefact known as the Villa of Oplontis.

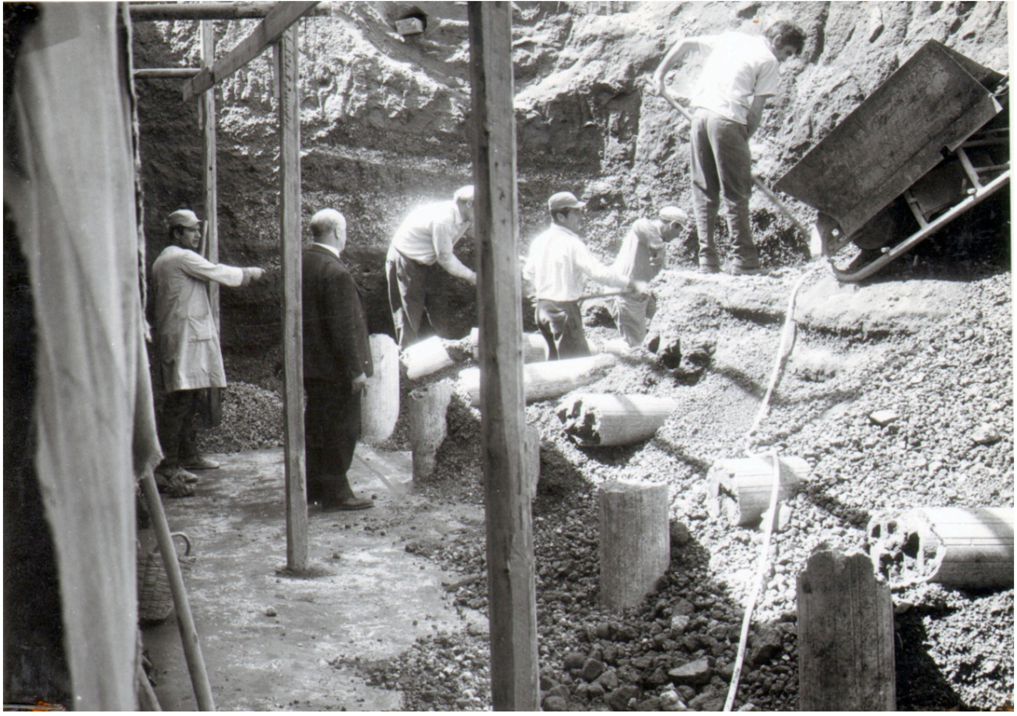


Figure 3. The discovery of porticus 33, circa 1967. Photo Soprintendenza Archeologica di Pompei D106

Archival research can greatly expand our knowledge—not only of reconstruction techniques but also of the Villa's decorative apparatus. In 1986, I became aware, through a drawing executed in 1971, that there once was a painting in the tympanum above the north alcove in room 11. After years of searching, in 2008 the archivist at the Superintendency of Pompeii found 30 black-and-white photographs taken in 1967 (Figure 4). The photograph clearly shows the subject of that painting, illegible today. It is a beautiful rendition of a harbour surrounded by turreted walls, populated by graceful figures. Using PhotoShop, Martin Blazeby of KVL has been able to create an accurate colour restoration of this important painting, allowing us to integrate it into the room's decorative scheme. It is worth mentioning that this painting is the only existing example of a landscape used as a tympanum decoration in this period, ca. 50 B.C. and that it has attracted much scholarly attention.

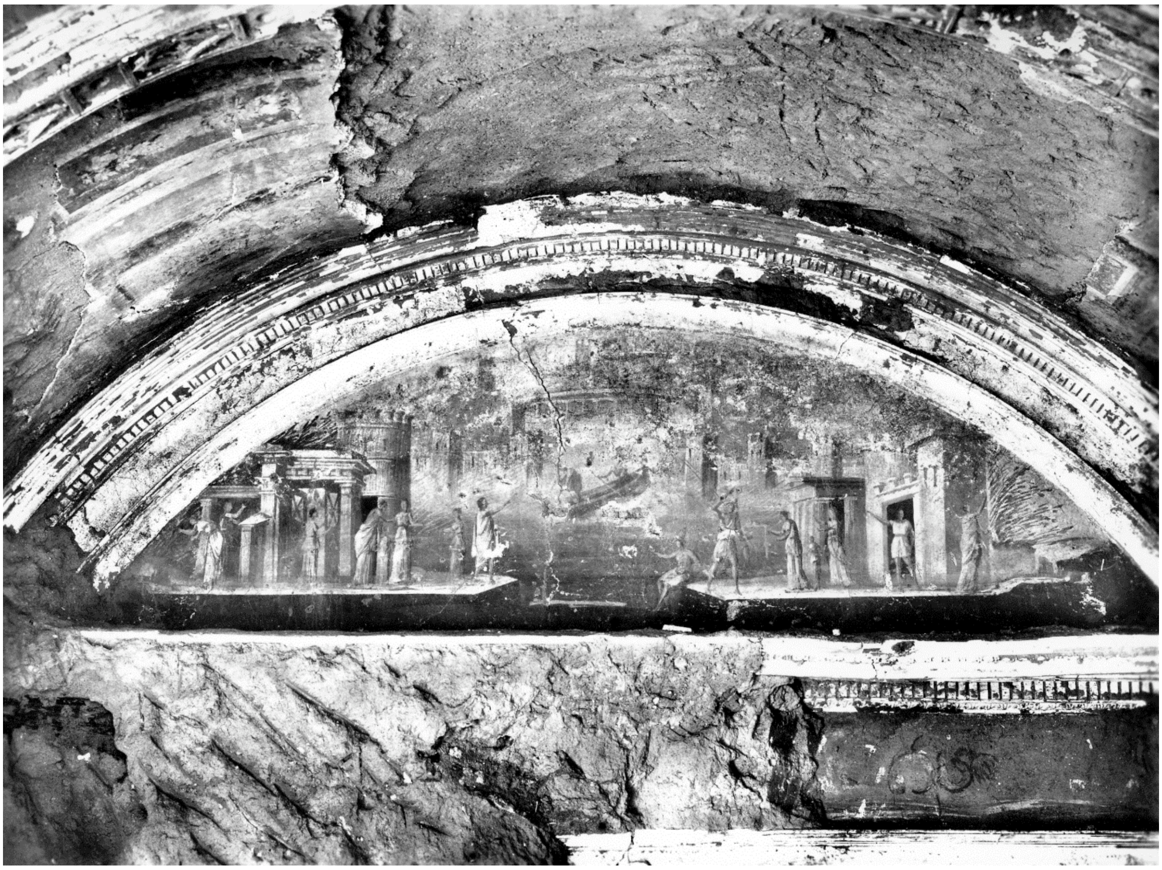


Figure 4. Room 11, north wall, tympanum, circa 1967. Photo Soprintendenza Archeologica di Pompei B1244

We get a sense of the chaotic situation the workers faced in another early archival photo, when excavations have got down to the upper third of the standing walls (Figure 5). We are looking at the upper left section of the east wall of room 15. But perhaps the most dramatic photograph is this one, of oecus 15. It reminds us how violent the cataclysm was, and that violent earthquakes accompanied the eruption. Two sides of the same wall appear in the photograph. We see the portion of the wall of oecus 15 that is still standing. We can recognize the image of the tragic mask and just the head of the famous Oplontis peacock sitting on a ledge. But at the bottom of the photograph we can see the image of a goddess in her circular shrine from the opposite side of this wall: a piece of the west wall of triclinium 14 overturned by the cataclysm.

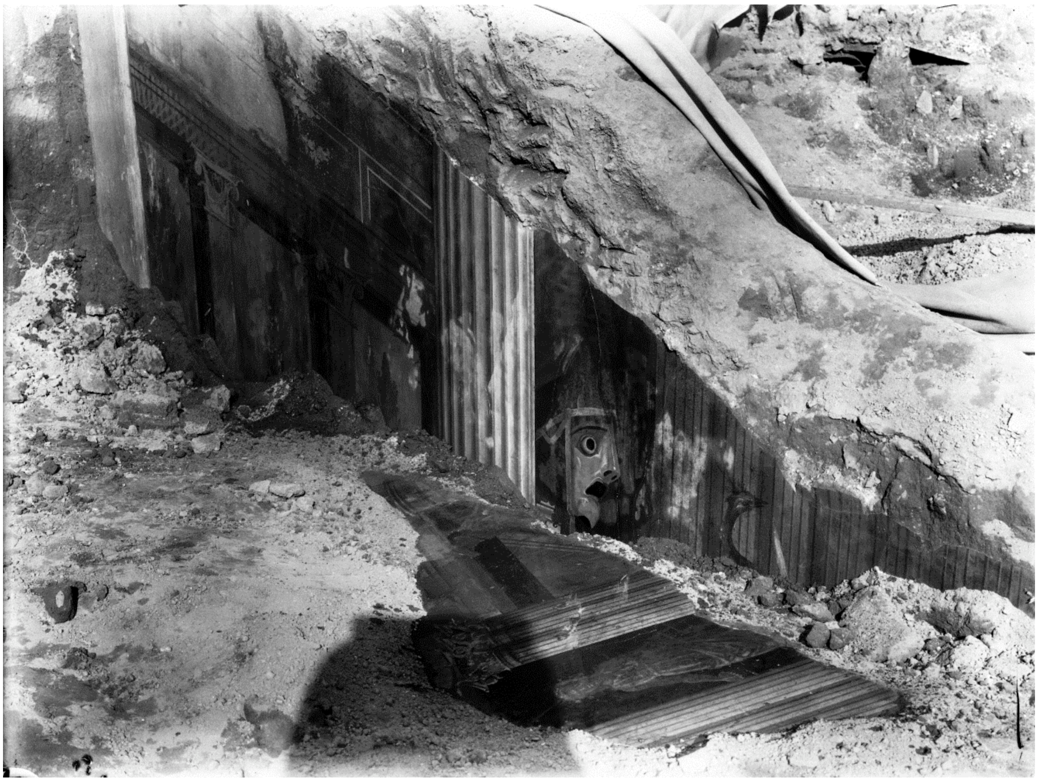


Figure 5. Oecus 15 and toppled wall of 14. Photo Soprintendenza Archeologica di Pompei B1284

Fortunately, a conscientious draftsperson, Ciro Iorio, recorded these fragments before they were lost forever in the failed attempts to recompose them and reattach them to the reconstructed wall (Figure 6). Today the entire upper left quadrant of Iorio's drawing is lost, including the frieze decoration consisting of shields and armour, the ornate capital at the top of the large pier that runs from floor to frieze, as well as the shield image, the so-called *imago clipeata*, to the left of the goddess in her circular shrine.

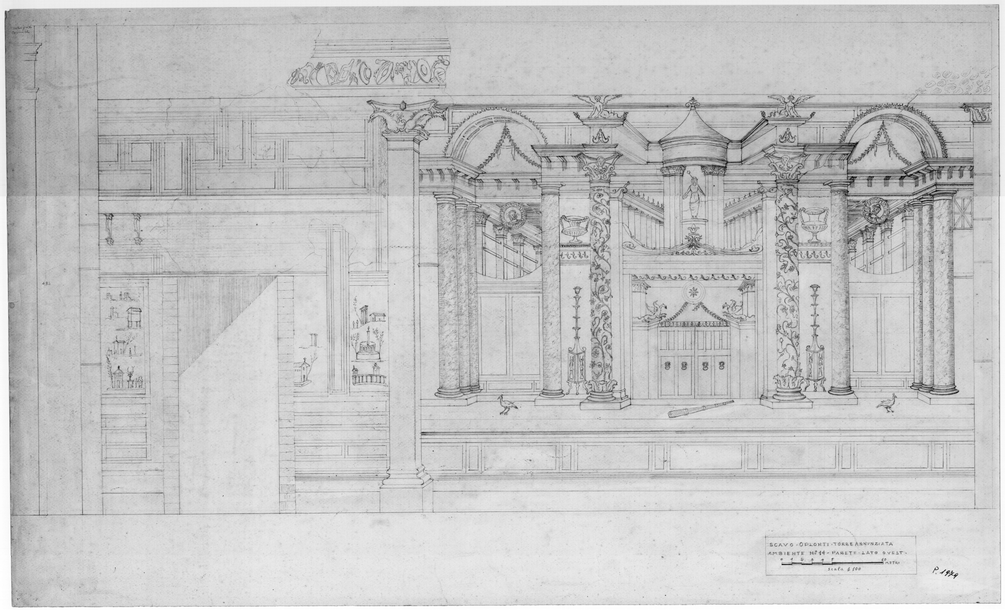

SCAVO - OPLONTI - TORRE ANNUNZIATA
AMBIENTE N° 14 - PARETE - LATO OVEST -
0 1 2 3 4 5 10 METRI
scala 1:100
P. 1979

Figure 6. Triclinium 14, west wall. Graphic reconstruction by Ciro Iorio, 1968. Drawing Soprintendenza Archeologica di Pompei P1979

We have used these archival photographs and drawings to create reconstructions of the original states of these and other decorations. Thanks to the work of Blazeby, the digital model allows the viewer to switch back and forth between actual-state and reconstructed views of the frescoes in many of the rooms of the villa. The digital model also allows us to distinguish separate phases of decoration in several rooms. In A.D. 45, the owners decided to convert room 8 from the hot room of a bath to a dining and reception space, the back wall had to be partially demolished. The owner then got an artist to imitate quite carefully the original—and by now, old fashioned— wall scheme. He or she did not want lose this period-style room—painted in A.D. 10—when the room was remodelled forty years later. By far the most dramatic find from our excavations is a piece of the demolished part of this very wall. It is a piece of the original frieze of the Third-Style decoration of A.D. 10 found in Oplontis Project trench OP3, about 100 metres away and buried along with lots of other plaster fragments at a depth of 1.5 metres (Clarke and Thomas 2009). Once again, this is the kind of scholarly information that we can imbed in the 3D model and explain clearly through the database documenting our excavations.

By coupling archival photographs with the 3D model we can also recover the original architectural configurations of spaces lost in the hasty and sometimes inept modern reconstruction process. In a photograph of the excavation of room 23 around 1967, we see workmen exploring the standing the north and west walls (Figure 7). They are clearing the rubble as best they can. They have put cloths weighted down by stones on the tops of the walls, and there are lots of wooden poles at the ready to build a temporary tin roof. Two important features are about to be lost in an undocumented but certain collapse: the remains of a tympanum on the north wall and the beginning of another on the west. In other words, the original covering of this room was a suspended cross-vaulted ceiling. The room got restored with a flat ceiling in reinforced concrete, no trace of either tympanum to be found. In its original form it would have resembled a well-preserved cross-vaulted room of the same period (ca. 50 B.C.) in the House of Ceres at Pompeii. Given the information provided by the archival photograph and the House of Ceres, we have with certainty restored properly the cross-vaulted ceiling of this beautiful space. This reconstruction demonstrates how archival photographs combined with archaeological research can create new knowledge.

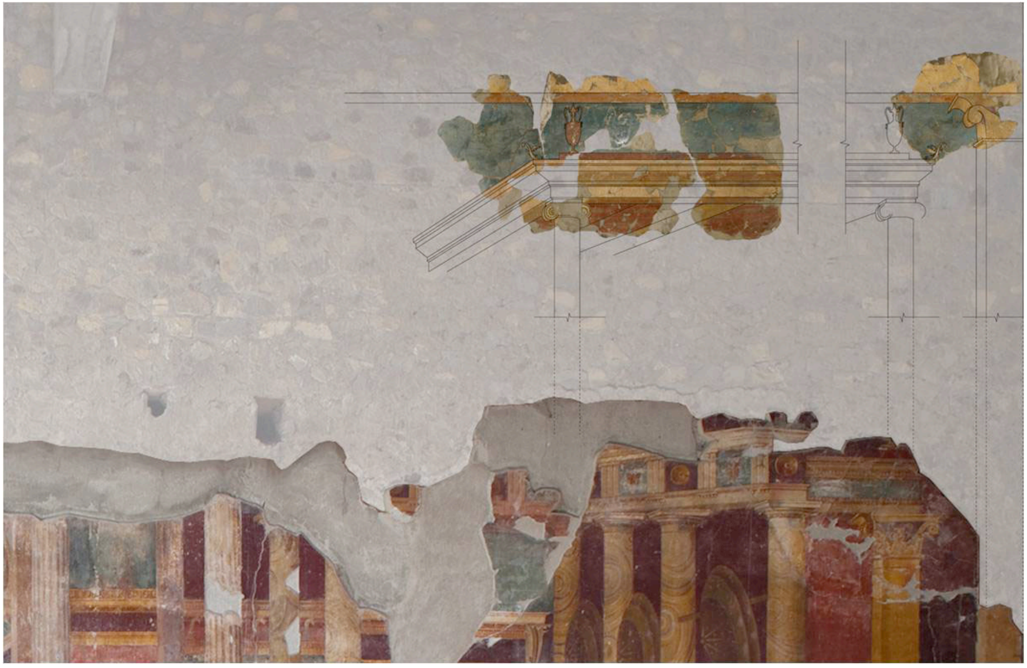


Figure 7. Atrium 5, east wall. Reconstruction of Ionic upper order. Timothy Liddell

In addition to filling in losses and correcting inept reconstructions, the digital model can provide a home for the hundreds of orphaned fragments in the villa. When workmen found a detached fragment of wall painting, they placed it face down on a table. They cut down the ancient plaster backing to create a regular surface; then they put down galvanized steel wire for reinforcement. They also formed this wire into loops so that they could hook the fragment back onto the reconstructed wall. They then built up modern cement around the reinforcing wire. Once the fragments from a room were consolidated in this way, they would cement them into the reconstructed wall—presumably in the right position.

Archival photographs from the early 1970s document many of these consolidated fragments that were never integrated into their original walls. Particularly in the atrium (5), many fragments of its decoration turned up after the hasty reconstruction, in reinforced concrete, of its roof—an imitation of the typical Roman impluviate type.

We located most of these fragments belonging to the Second-Style scheme of the atrium in a storage area of the Villa. One of our architects, Timothy Liddell, worked with the puzzle. Using Photoshop and Illustrator, he was able to place a number of these orphaned fragments into a plausible scheme that fits with the architectural perspectives of the standing wall (Figure 8). Since the only column-capital among the atrium fragments is Ionic, Liddell hypothesized an Ionic upper order. The pieces fit quite well; however, it is impossible to calculate the exact height of the lost Ionic columns. Liddell's reconstruction is deliberately quite conservative, and for this reason only some of our orphaned fragments have found a home. A full, intelligently-imaginative reconstruction of the decoration, recently completed by Blazeby, takes Liddell's Ionic architrave and elaborates an upper story (Figure 9). These two reconstructions come from combining two kinds of archives: photographs and the actual fragments found in deplorable condition storage in one of the rooms of the Villa that had become a rubbish heap.

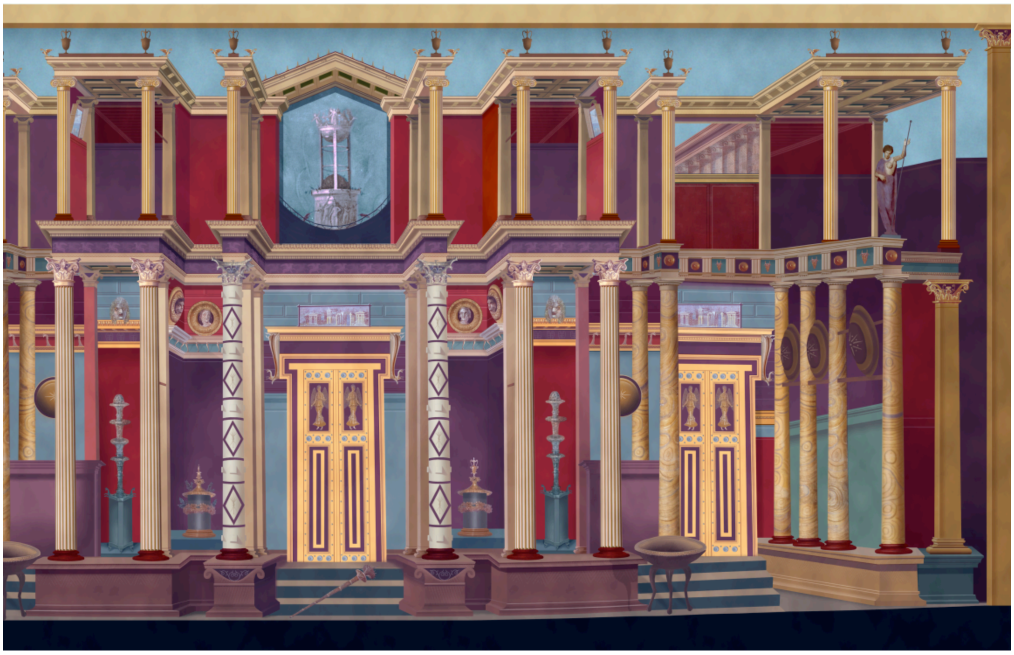


Figure 8. Atrium 5. Reconstruction of east wall. Martin Blazeby

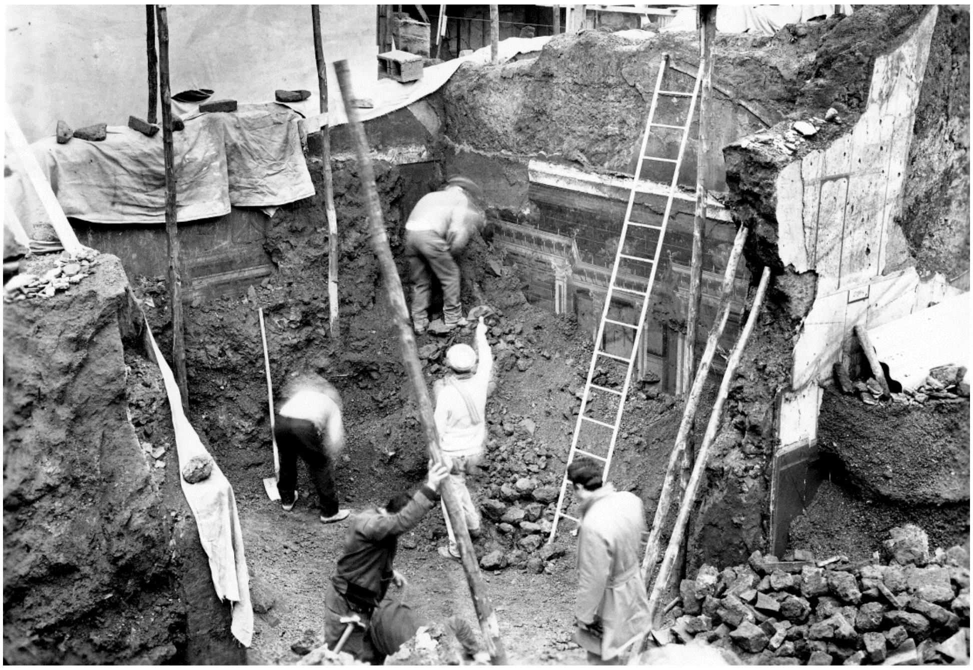


Figure 9. Room 23, excavation of north and west walls, circa 1967. Photo Soprintendenza Archeologica di Pompei, De Franciscis 945

Work continues on the many other fragments that constitute a physical archive of the Villa's decoration. Recently we found numerous fragments of a Fourth-Style scheme that was partially consolidated but never fully restored. These fragments were extremely delicate, not only because the plaster was so friable but also because the painters applied the costly cinnabar red pigment as like a watercolour wash rather than allowing the pigment to carbonate in true fresco fashion. After careful study by team members Erin Anderson and Zoe Schofield, these fragments turn out to be part of the lost ceiling of the room 8—the same room as that where we found a match for our excavated frieze fragment. Back in 1967, excavators made only a half-hearted attempt to restore this ceiling. There are many other diagnostics that make it certain that this ceiling fragment, after wasting away for 44 years in a dustheap, has finally found a home. This major discovery restores a ceiling to one of the most important rooms of the Villa. Another group of fragments restores the lost ceiling of porticus 60. As work continues, we hope to find more pieces of this decoration as we wade through the remaining orphaned fragments, and we hope to find a home for them—at least in the digital model.

Another invisible element of the Villa is its sculpture. The pieces found in situ, in the north garden (56) and along the eastern side of the 61 m. swimming pool, have never been exhibited at the archaeological site. We hope to put the sculpture found in these gardens back into place, after languishing for forty years in the storerooms. Now and then a few pieces get pulled out of storage to illustrate a theme of an exhibition. We intend to put them into the digital model, so that a viewer will be able to see them from various vantage points in the villa, walk around them, and contemplate them as an ancient Roman would have done.

The model will not only be a perfect medium for studying the sculptures, it will also allow us to demonstrate, finally, just how Villa A fit into the ancient landscape. The modern coastline has changed dramatically over the centuries. The masses of materials blasted from the cone of Vesuvius in A.D. 79, as well as the many subsequent eruptions, have changed its shape dramatically. Over the past years, geologist Giovanni di Maio conducted a series of cores, most of them going 15 metres deep, to determine the shape of the ancient coastline. Along it, archaeologists have noted the foundations of several ancient villas in addition to Villa A. In di Maio's aerial view, he has marked the remains of these villas with red crosslets and Villa A with a circle (Figure 10). Today, Villa A lies about half a kilometre inland, at the depth of 8.5 metres below the level of the modern city. In a second view, we see the outline of the ancient coast, the position of Villa A marked with a crosslet (Figure 11). In 2009-2010 the Oplontis Project commissioned di Maio to conduct cores around the Villa to determine just what it stood on. The result is rather dramatic as the section reveals (Figure 12). The Villa was not on flat land, not even on sloping land. It stood on a cliff 13 metres above the sea. It would have commanded a dramatic view. We have already explored this ancient panorama as seen from various rooms overlooking the sea using Google Earth and other GIS imaging tools.

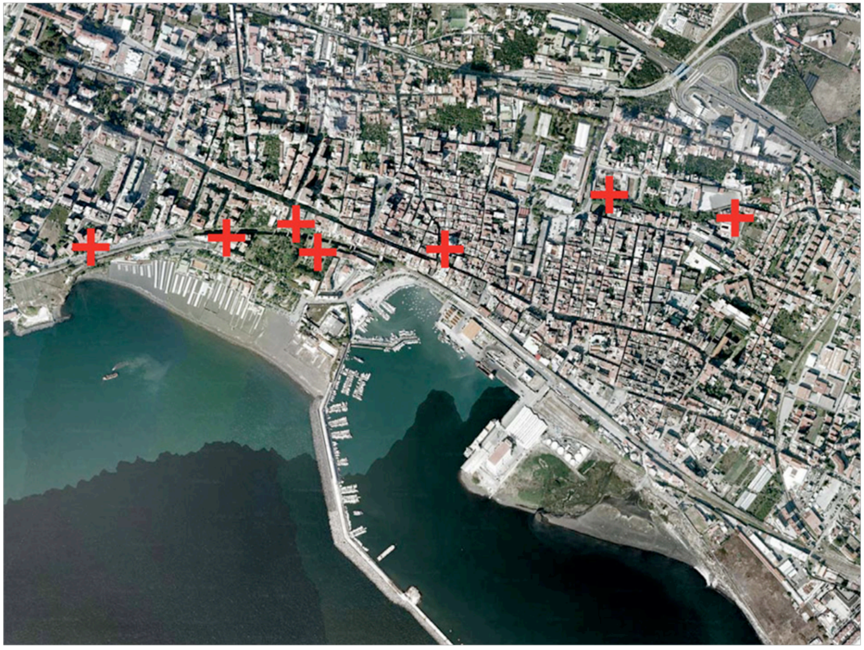


Figure 10. Modern coastline at Torre Annunziata, with sites of ancient Roman villas marked. Giovanni di Maio

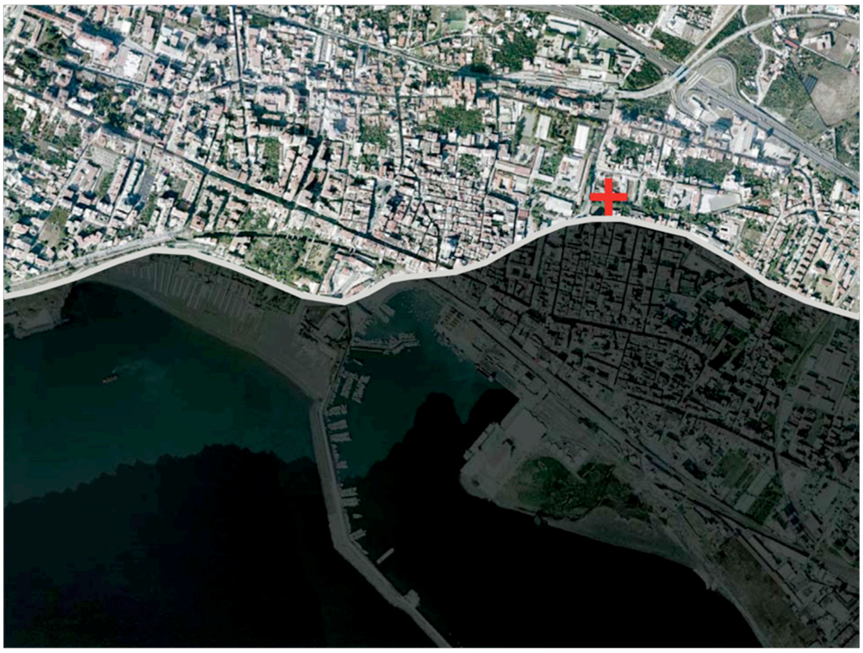


Figure 11. Reconstruction of the ancient coastline of Torre Annunziata, A.D. 79 with position of Villa A marked. Giovanni di Maio

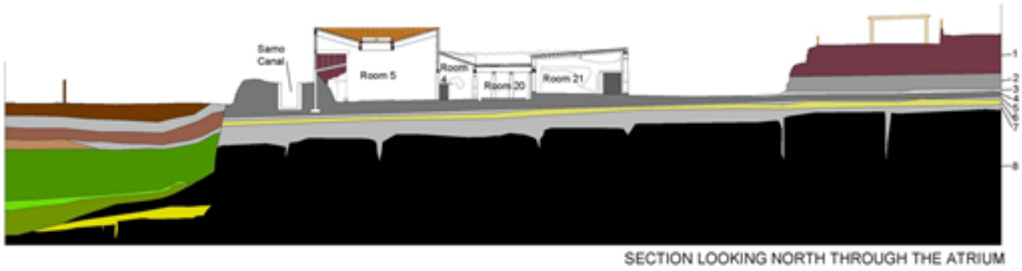


Figure 12. Section of Villa, south to north, at atrium. Giovanni di Maio

Conclusion

The excavations at the so-called Villa of Poppaea stand at a crossroads in the long history of excavations in the region buried by Vesuvius. They aimed, as Amedeo Maiuri had at Herculaneum, to make the Villa into a living museum that the public could visit. This meant creating a new building that looked ancient. Walls had to be rebuilt and colonnades had to be reconstructed to support modern concrete beams and new tile roofs. In the process much of what made the villa so attractive in antiquity had to be sacrificed. The sculpture had to be removed to safekeeping, as well as all objects subject to vandalism or thievery. Archives, like precious excavation daybooks, disappeared. When funds dried up, restorers gave up on the fragments of painting that did not easily fit back on the walls. The gardens never got restored properly, despite the wealth of evidence discovered by the American garden archaeologist, Wilhelmina Jashemski.

We can now, through virtual means, put the pieces of this puzzle back together. As we have seen, the idea of the archive of an excavation is changing from that of a catalogue of objects classified according to their materials into an interwoven and—indeed—an interactive experience. The ideal archive is one that allows us, and future generations, to find material easily electronically, and to study it in the context of the place where it originally functioned. We hope to facilitate just that kind of retrieval and study with the Oplontis Project, linking our database with the interactive 3D model, putting the fragments back in place, showing what came from our trenches and cores, and making this information—as well as the four-volume publication—available for free on the web. In this way, if and when excavation of the Villa continues—and I estimate that another 40 percent of the structure still lies buried—scholars of the future will be in much better shape than we were when we began our study. They will have at their fingertips all of the information they need to complete the study of this rich and complex Villa.

References

American Council of Learned Societies (ACLS), Humanities E-Book Series. Retrieved from http://www.humanitiesebook.org/.

Oplontis Project website. (2011), from http://www.oplontisproject.org/

Clarke, J. R., & Thomas, M. L. (2009). Evidence of demolition and remodeling at Villa A at Oplontis (Villa of Poppaea) after A.D. 45. *Journal of Roman Archaeology, 22*, 201-209.

Ling, R. (2005). *The Insula of the Menander at Pompeii, vol. 2. The Decorations*. Oxford: Clarendon Press.

Strocka, M. V. (Ed.). (1991-.). *Häuser in Pompeji*. 12 vols. Munich: Hirrmer Verlag.

The ISDA Tools: Preserving 3D Digital Content

Kenton McHenry, Rob Kooper, Luigi Marini, Michael Ondrejcek

National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, 1205 W. Clark St., Room 1008, Urbana, IL 61801 USA

Abstract. In collaboration with the U.S. National Archives and Records Administration Division of Applied Research; the Image, Spatial, and Data Analysis group at NCSA has developed a number of tools to aid in the preservation of digital records. In this paper we present these tools in the context of preserving 3D digital files. Tools such as the Conversion Software Registry, Software Servers, Polyglot, 3D Utilities, and Versus provide users with a scalable means of discovering and carrying out large file migration tasks in a manner that can take into account and quantify the unavoidable information loss that occurs when moving from one format to another. We present each of these tools and describe how they can be used.

Introduction

There are many different file formats to store any given content type. This is especially true in the case of 3D content where it seems that nearly every vendor of 3D software tends to come up with their own unique file format (Figure 1). In our evaluation of a number of popular 3D software packages we have documented over 144 different file formats (McHenry et al. 2011). Having many different formats for the same type of data is a problem for a couple of reasons. The first reason is a matter of accessibility in that having many formats makes it difficult to share data. If a particular format is obscure or only used by one particular software application, then content created in that application might be difficult to share with users who do not have that software. The second reason is a matter of preservation. If a software vendor uses a proprietary format and does not make the format's specification open, then if that vendor were to ever go out of business, any content stored within that format could be potentially locked away forever. This latter situation has been known to occur.

One might argue that in an ideal world there would only be one file format for each type of content. Having one format, a format that has a standardized open specification, would make archiving that content much easier in that even if a viewer no longer exists in the future for that format, a new viewer could be created based on the available specification. Determining what this format could be is a problem. Even once this format is determined, converting from all the formats available today to that one format is also a problem.

As soon as the need for file format conversion arises we must begin to consider information loss. In Figure 2 we list several 3D formats along the types of information they each support. As shown here, 3D files store a variety of attributes from geometry, to appearance, to scene properties and animation. Not all file formats support all attributes. Even within individual attributes, information can be stored differently. For example the geometry of a model is often represented as either a faceted mesh, as parametric surfaces, as a boundary representation, or as constructive solid geometry. Converting from a format that supports one representation to one that supports another representation requires a conversion of the content itself. Some of these conversions are possible. For example B-Rep[1] surfaces can be transformed to faceted meshes by a process called tessellation. Doing this will alias an otherwise continuous surface and depending on the sampling drastically change the file size. The tessellation process will result in a loss of information in that the inverse process, going from a faceted mesh to a B-Rep, will not result in the original content. In fact this inverse conversion is not trivial at all.

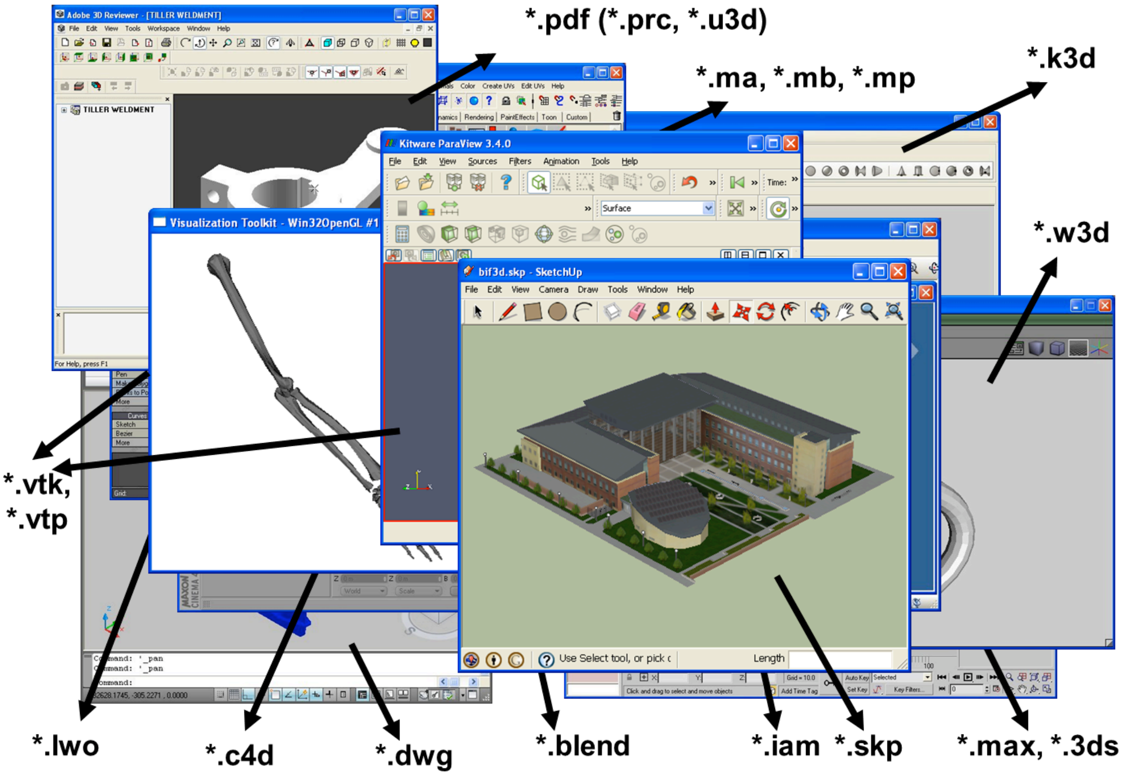


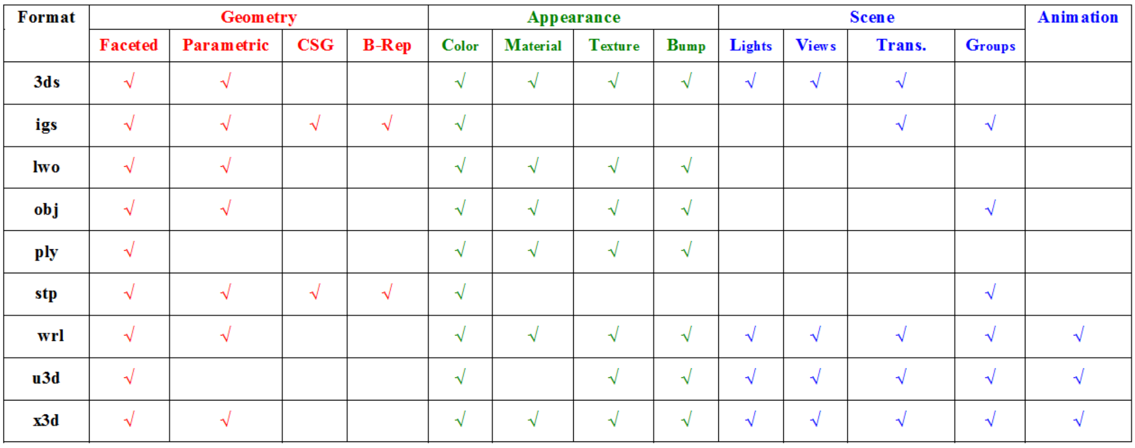
Figure 1. Software vendors have a tendency to introduce new formats for the content that their software saves. This is especially true in the case of 3D content which has a large number of available formats today.

Keeping in mind that conversions will almost always result in some sort of information loss we can then begin the process of determining what is an optimal file format for long term preservation.

We define this format to be one that is standardized/open and results in as little information loss as possible when converted to from all other available formats. Such a format would have the best chance of keeping the bulk of an archive’s data accessible in the future.

Below we present a number of tools we have developed towards the end goal of empirically evaluating which format is optimal for long term preservation for a given content type. We present these tools individually as they possess useful qualities in their own right, separate from the overall goal of choosing an optimal file format.

[1] Boundary Representation



Format	Geometry				Appearance				Scene				Animation
	Faceted	Parametric	CSG	B-Rep	Color	Material	Texture	Bump	Lights	Views	Trans.	Groups	
3ds	√	√			√	√	√	√	√	√	√		
igs	√	√	√	√	√						√	√	
lwo	√	√			√	√	√	√					
obj	√	√			√	√	√	√				√	
ply	√				√	√	√	√					
stp	√	√	√	√	√							√	
wrl	√	√			√	√	√	√	√	√	√	√	√
u3d	√				√		√	√	√	√	√	√	√
x3d	√	√			√	√	√	√	√	√	√	√	√

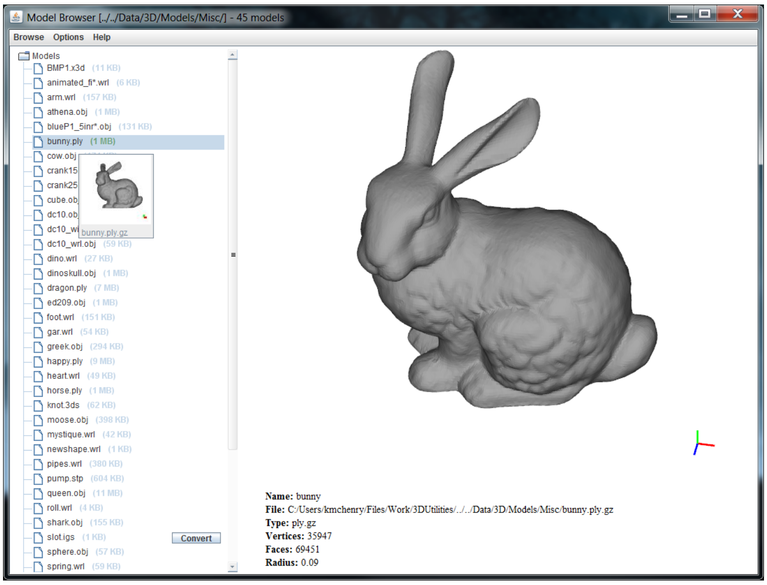
] t
attributes stored within 3D files are geometry, appearance, and scene structure. Each attribute can also be stored in a number of ways. Converting between formats that do not support the same type of information will result in some sort of information loss.

3D Utilities

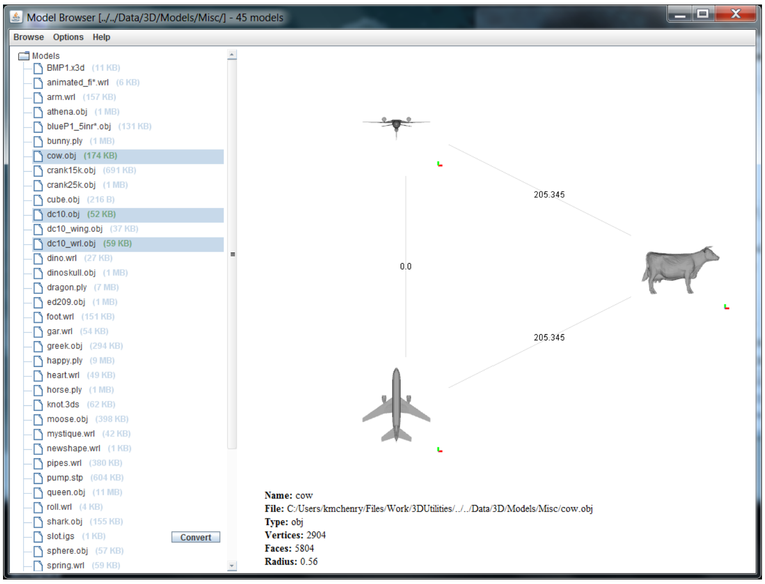
Our 3D Utilities are both a library and a collection of tools written in Java, created for the purpose of accessing content within various 3D file formats. Like the NCSA Portfolio project before it, 3D Utilities provides a number of 3D file loaders. Unlike Portfolio that was built on Java3D and loaded content in a Java3D scene graph, 3D Utilities takes a far simpler and less restrictive approach. 3D Utilities uses an extremely simple polygonal mesh representation for all its 3D content. File loaders for various formats are created so as to parse a 3D file type and load its content into a polygonal mesh. If a file format does not store its 3D geometry as a mesh, then it is up to the loader to convert between representations. This library of loaders can be used by any java application to load 3D content from a file for the purpose of rendering or manipulating it through the mesh data structure (which is nothing more than a list of vertices and faces connecting them).

In addition to file loaders, the 3D Utilities library also contains a library of mesh signatures. These mesh signatures act as a hash allowing one to compare two different 3D models and retrieve the most similar instance from a collection of 3D models. We have signatures implemented based on vertices statistics, polygonal face surface area (Brunnermeier & Martin (1999), light fields (Chen et al. 2003), and spin images (Johnson & Hebert 1999). Each signature allows for different aspects of the models to be considered during a comparison and each is best suited under differing situations. These signatures can be used within Java code to compare two meshes loaded using the library of loaders.

3D Utilities also contains several tools: ModelViewer, ModelViewerApplet, ModelBrowser, and ModelConverter. The ModelViewer is both a class extending a JPanel, which can be used within code to display 3D content, and a standalone application to view content within 3D files. The ModelViewerApplet is an applet version of the ModelViewer allowing 3D content to be viewable from within a web browser. The ModelBrowser tool uses ModelViewers to provide a convenient way of viewing all 3D files under a given file system directory. As shown in Figure 3 (top) all 3D files found under a user specified directory are shown in the leftmost panel. From here a user can select one of the files to be displayed in the top right panel. This panel being an instance of a ModelViewer allows a user to use the mouse to change viewing directions and manipulate the model. In the bottom right panel, metadata of the shown 3D model is displayed. If multiple files are selected in the left pane, they are compared using a specified signature and shown simultaneously in the right pane, with distances along edges that connect them. This can be used as a means of visualizing how the various signatures emphasize different aspects of a 3D model during comparison (Figure 3 bottom). The last tool, the ModelConverter, utilizes the 3D Utilities file loaders to perform format conversions.



Name: bunny
File: C:/Users/kmchenry/Files/Work/3DUtilities/../../Data/3D/Models/Misc/bunny.ply.gz
Type: ply.gz
Vertices: 35947
Faces: 69451
Radius: 0.09



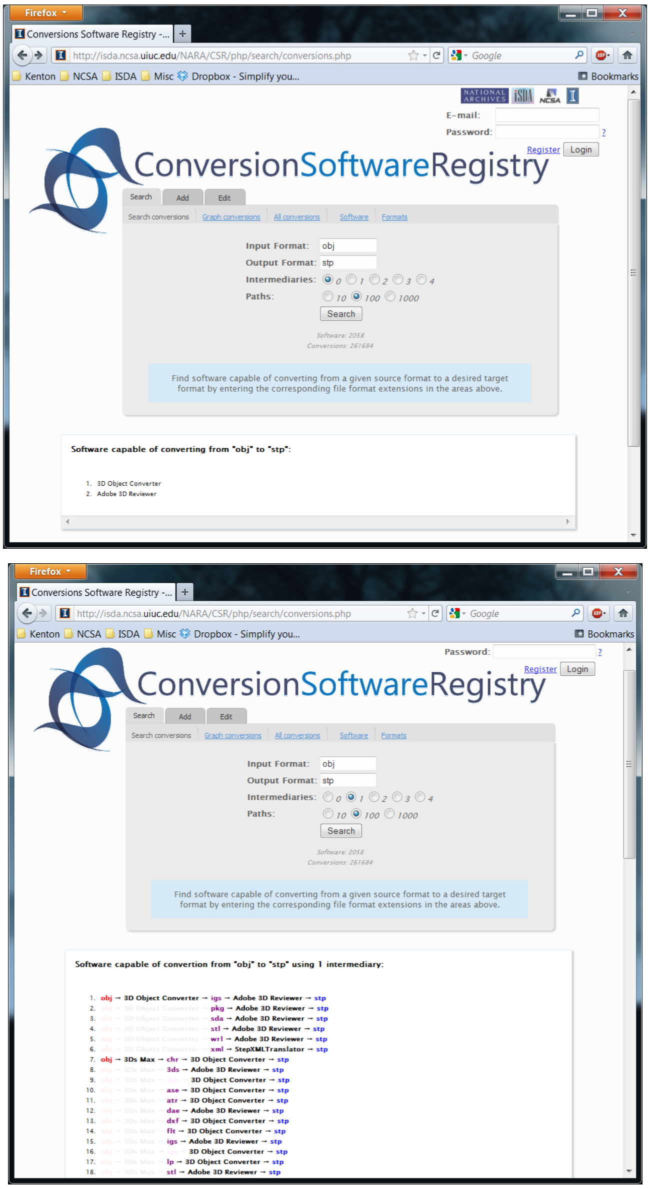
Name: cow
File: C:/Users/kmchenry/Files/Work/3DUtilities/../../Data/3D/Models/Misc/cow.obj
Type: obj
Vertices: 2904
Faces: 5804
Radius: 0.56

Figure 3. Top: The Model Browser tool included as part of the 3D Utilities allows one to view 3D files under a given directory and manipulate data within a number of different 3D file formats. Bottom: If a user selects multiple files from the left hand side, the 3D content from each file is displayed and compared to one another. Above 3 models are compared using the light fields measure (Chen et al. 2003), a measure that is invariant to rigid transformations such as rotations. Notice the two planes have a distance of 0 between them even though they are rotated differently.

As all loaders load/save content to/from a mesh representation, it is a simple matter to convert between formats by loading the content using one loader and saving it using another.

The Conversion Software Registry

As stated previously, vendors have a tendency to create new file formats. For the sake of some level of portability, applications often allow content to be imported/exported to and from a handful of other formats.



E-mail:
Password:

Input Format: obj
Output Format: stp
Intermediaries: ● 0 ○ 1 ○ 2 ○ 3 ○ 4
Paths: ○ 10 ● 100 ○ 1000
Search

Software: 2058
Conversions: 261684

Find software capable of converting from a given source format to a desired target format by entering the corresponding file format extensions in the areas above.

Software capable of converting from "obj" to "stp":

1. 3D Object Converter
2. Adobe 3D Reviewer

Password:

Input Format: obj
Output Format: stp
Intermediaries: ○ 0 ● 1 ○ 2 ○ 3 ○ 4
Paths: ○ 10 ● 100 ○ 1000
Search

Software: 2058
Conversions: 261684

Find software capable of converting from a given source format to a desired target format by entering the corresponding file format extensions in the areas above.

Software capable of convertion from "obj" to "stp" using 1 intermediary:

1. obj → 3D Object Converter → igs → Adobe 3D Reviewer → stp
2. obj → 3D Object Converter → pkg → Adobe 3D Reviewer → stp
3. obj → 3D Object Converter → sda → Adobe 3D Reviewer → stp
4. obj → 3D Object Converter → stl → Adobe 3D Reviewer → stp
5. obj → 3D Object Converter → wrl → Adobe 3D Reviewer → stp
6. obj → 3D Object Converter → xml → StepXMLTranslator → stp
7. obj → 3Ds Max → chr → 3D Object Converter → stp
8. obj → 3Ds Max → 3ds → Adobe 3D Reviewer → stp
9. obj → 3Ds Max → 3ds → 3D Object Converter → stp
10. obj → 3Ds Max → ase → 3D Object Converter → stp
11. obj → 3Ds Max → atr → 3D Object Converter → stp
12. obj → 3Ds Max → dae → Adobe 3D Reviewer → stp
13. obj → 3Ds Max → dxf → 3D Object Converter → stp
14. obj → 3Ds Max → flt → 3D Object Converter → stp
15. obj → 3Ds Max → igs → Adobe 3D Reviewer → stp
16. obj → 3Ds Max → igs → 3D Object Converter → stp
17. obj → 3Ds Max → lp → 3D Object Converter → stp
18. obj → 3Ds Max → stl → Adobe 3D Reviewer → stp

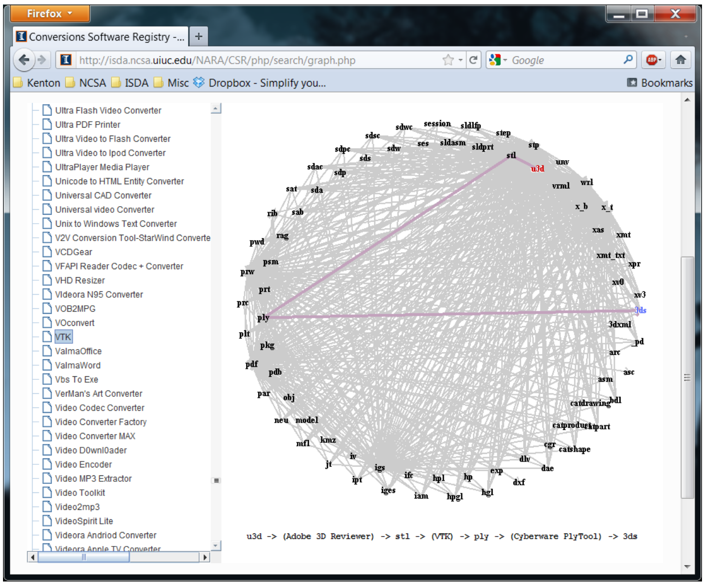
Figure 4. The Conversion Software Registry. Top: A user searching for software based on supported for input format *.obj and output format *.stp. Bottom: A user searching for chains of software supporting this same input and output format, however, this time with one allowed intermediary format.

This tendency allows many software applications to act as converters between different formats. As many formats are closed source and proprietary, the software of the vendor responsible for the format is often the best means of getting into and out of that particular format.

As many applications only support a small number of imports/exports, it is also very likely that many desirable conversions from specific source formats to target formats will not be supported. However, it is very possible that multiple applications chained together can carry out conversions though intermediary formats that are not directly supported by any one application.

This information, in terms of software inputs and outputs, is essential in order to identify software to carry out a specific conversion. Determining what software is needed for a particular conversion, however, can be difficult without direct access to the software (i.e. without purchasing the software). To aid users in this task we have created the Conversion Software Registry (CSR), an online database that indexes software based on the input and output formats they support (Ondrejcek et al. 2010). As shown in Figure 4 (top) a user can type in the extension of a source format and target format to get back a list of applications that directly supports that conversion. By clicking a radio button below the query box, a user can tell the system to consider conversions using one or more intermediary formats. As shown in Figure 4 (bottom) many more options are presented by doing this.

The CSR also allows users to search through the space of possible conversions graphically (Figure 5). Applications selected in the left pane are represented in the right pane as an input/output graph. This graph has at its vertices file formats and directed edges connecting a pair of formats to represent an application capable of carrying out that particular conversion. A user can search for a conversion path between formats by using their mouse to select a source and target file format. An un-weighted shortest path algorithm is then used to find a conversion path with the least number of intermediary formats.



Ultra Flash Video Converter
Ultra PDF Printer
Ultra Video to Flash Converter
Ultra Video to Ipod Converter
UltraPlayer Media Player
Unicode to HTML Entity Converter
Universal CAD Converter
Universal video Converter
Unix to Windows Text Converter
V2V Conversion Tool-StarWind Converte
VCDGear
VFAPI Reader Codec + Converter
VHD Resizer
VIdeora N95 Converter
VOB2MPG
VOconvert
VTK
ValmaOffice
ValmaWord
Vbs To Exe
VerMan's Art Converter
Video Codec Converter
Video Converter Factory
Video Converter MAX
Video D0wnl0ader
Video Encoder
Video MP3 Extractor
Video Toolkit
Video2mp3
VideoSpirit Lite
Videora Andriod Converter
Videora Apple TV Converter

```
u3d -> (Adobe 3D Reviewer) -> stl -> (VTK) -> ply -> (Cyberware PlyTool) -> 3ds
```

Figure 5. A user can search the data within the CSR graphically. After selecting a number of applications to consider on the left, the user can use the mouse to select a source and target file format. If a means of converting between the source and target file format exists the shortest chain of software capable of carrying it out will be displayed.

Other types of supported queries include: finding all the formats reachable from a given source format, finding all the formats that can reach a particular target format, and finding a set of file formats that can be commonly reached by a selected set of source formats.

Software Servers

Ideally we would like to be able to support format conversions within code (e.g. Java). However, because of the many closed propriety formats in existence, this is not at all feasible. As pointed out in the previous section, nevertheless, there is software available to carry out many conversions. A Software Server is a tool that attempts to bridge this gap between need and availability (McHenry et al. 2011, 2010; McHenry et al. 2009b). A Software Server running on a system allows functionality within locally installed software to be shared in a matter that is analogous to how Windows allows one to share a folder. To share functionality within software (e.g. the open and save operations within a particular program), one needs only right click on the shortcut of the application and select "Share Functionality" from the menu (Figure 6).

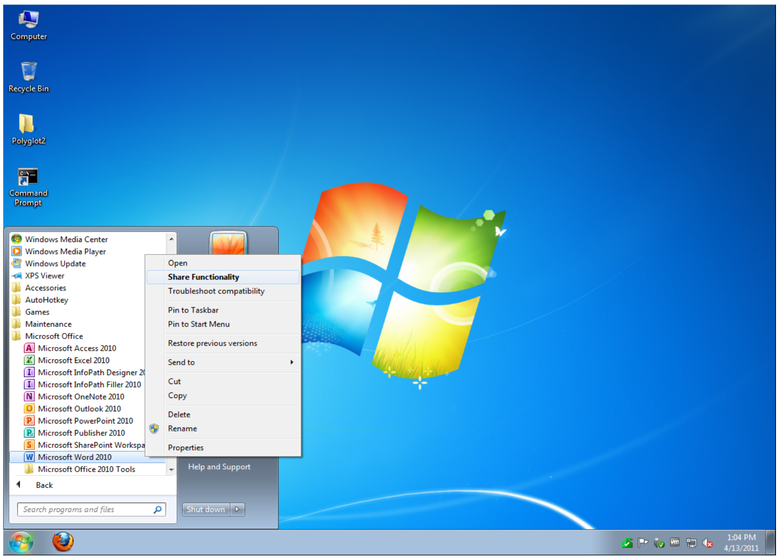


Figure 6. Software functionality can be shared in a manner analogous to how folders are shared in the Windows operating system. Access to operations within software can be shared simply by right clicking on them and selecting share from the menu.

Once this is done information obtained from the shortcut is used to query the Conversion Software Registry, which serves as a repository for wrapper scripts that are capable of automating functionality within command line and graphical interface driven software. These scripts can be written in any text based scripting language. We tend to use AutoHotKey[1] and Sikuli (Yeh et al. 2009) as they allow for the scripting of graphical interfaces (making a large amount of software accessible from the Software Server). Wrapper scripts associated with the needed software are downloaded and configured to run on the local environment.

Once a Software Server is running shared software, functionality can be accessed by simply pointing a web browser to the address of the hosting machine (Figure 7 top). When this is done, a user is presented with a page that mimics modern file managers with a number of icons representing available software. When a user clicks on a particular piece of software, they are presented with a form that allows them to call the remotely shared functionality from the browser (Figure 7 bottom). Tasks always take the form of a quadruple of the form: software, task, output format, input file. Though the form presents

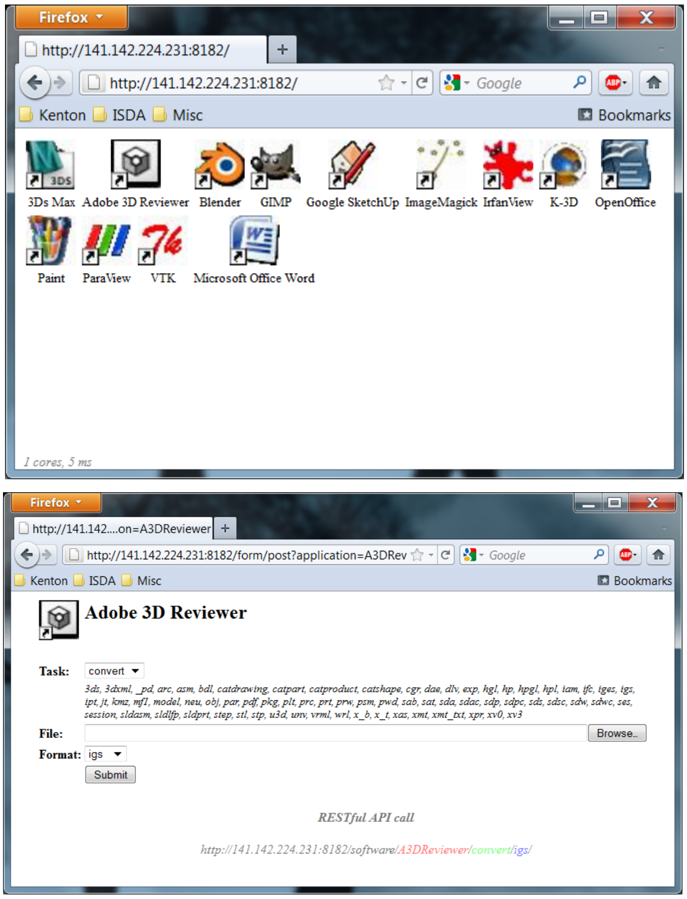
[1] http://www.autohotkey.com

this in an easy to use graphical manner, the strength of the Software Server is in the RESTful interface[1] it provides to carry out these tasks.

Specifically the same task can be carried out by posting a file to a URL on the host in the form of:

`http://<host>:8182/software/<Software>/<Task>/<Output>/`

where the final element of the quadruple, the input file, is the posted file. This simple, consistent, widely accessible interface to software allows software functionality to be used within new code as if it were a function within a library. Any language that supports accessing URLs can access this functionality, possibly wrapping it to look like a native library.



Adobe 3D Reviewer

Task: convert

3ds, 3dxml, _pd, arc, asm, bdl, catdrawing, catpart, catproduct, catshape, cgr, dae, dlv, exp, hgl, hp, hpgl, hpl, iam, ifc, iges, igs, ipt, jt, kmz, mf1, model, neu, obj, par, pdf, pkg, plt, prc, prt, prw, psm, pwd, sab, sat, sda, sdac, sdp, sdpc, sds, sdsc, sdw, sdwc, ses, session, sldasm, sldlfp, sldprt, step, stl, stp, u3d, unv, vrml, wrl, x_b, x_t, xas, xmt, xmt_txt, xpr, xv0, xv3

File:

Format: igs

RESTful API call

http://141.142.224.231:8182/software/A3DReviewer/convert/igs/

Figure 7. Shared software functionality can be accessed from a browser by accessing the host address of the machine running the software server. Top: If accessed directly from a browser, the software is presented in a manner that resembles a file manager from a modern OS with software represented by icons. Bottom: When an icon is clicked on, the user is presented with a form that allows them to access the software functionality.

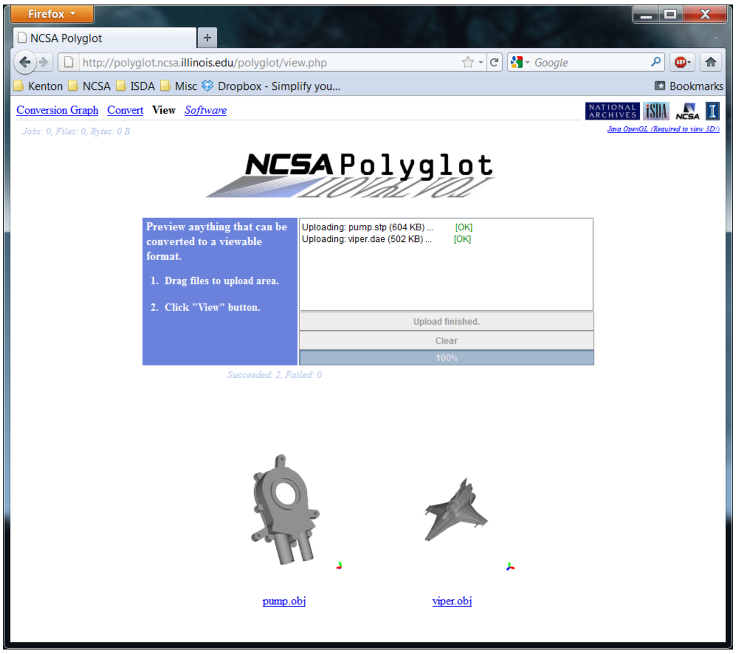
Polyglot

NCSA Polyglot (McHenry et al. 2009a) is a conversion service built on top of the previously described Software Servers. When Software Servers come online they begin to broadcast their existence. A Polyglot instance listens for these broadcasts. When a new Software Server is found, Polyglot will query it for all shared software functionality, looking for software with available input and output operations to carry out conversions. From this information Polyglot constructs an input/output graph. This graph is searched

[1] http://en.wikipedia.org/wiki/Representational_state_transfer

in order to carry out conversions across chains of software based on a given input format and desired output format.

Polyglot can be accessed through either a Java API or a web interface which allows a user to drag and drop a number of files from their local machine, select an output format, and carry out the conversions with the results being made available for download (Figure **8**).



Preview anything that can be converted to a viewable format.

1. Drag files to upload area.
2. Click "View" button.

Uploading: pump.stp (604 KB) ... [OK]
Uploading: viper.dae (502 KB) ... [OK]

Succeeded: 2, Failed: 0

Figure 8. A web interface to NCSA Polyglot. A user can drag and drop files to the large area in the middle. In this particular case the user does not select the desired output format as it is set automatically to one that can be viewed within the browser. When the user presses the "View" button, a conversion path is executed across the underlying Software Servers to reach the desired target format. When completed, the results are displayed in the area below.

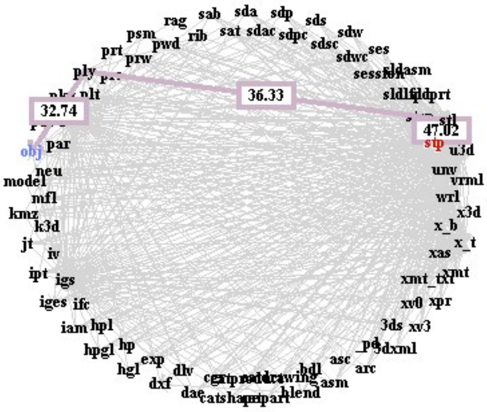


Figure 9. A weighted input/output graph, where the weights represent information loss obtained using Versus on a sample set. A shortest weighted path algorithm can be used to find conversion paths that result in the least amount of overall information loss.

Versus

Versus is a framework/library of content-to-content comparison measures supporting raster images, vector graphics, 3D models (via the signatures from our 3D Utilities), and documents. These measures are used by Polyglot as a means of estimating the information loss incurred by converting from one format to another through a piece of software. Information loss is evaluated in Polyglot by using a sample set of files in a format that can be directly loaded by one of the loaders in our 3D Utilities library. These files are converted to every reachable format within an input/output graph and then converted back to the original. We are forced to execute through these A-B-A' conversion paths in order to compare the file's contents before and after the conversion. Each before and after file is compared using a set measure from Versus, and the returned similarity is averaged along each edge representing software in the graph (Figure 9). This weighted graph can be used by Polyglot to identify conversions paths with minimal information loss. In addition, this weighted graph can be used to derive an answer to the question of which format would be best suited for long term preservation (as described earlier).

Conclusion

The ISDA tools are a collection of libraries and software constructed for the purpose of providing solutions to problems in digital preservation. We have presented five of these tools: 3D Utilities, the Conversion Software Registry, Software Servers, Polyglot, and Versus. These and others are available from our site at http://isda.ncsa.illinois.edu as free open source software.

Acknowledgments. This research has been funded through the National Science Foundation Cooperative Agreement NSF OCI 05-25308 and Cooperative Support Agreement NSF OCI 05-04064 by the National Archives and Records Administration (NARA).

References

Brunnermeier, S., & Martin, S. (1999). Interoperability cost analysis of the U.S. automotive supply chain: RTI International Research Publications.

Chen, D., Tian, X., Shen, Y., & Ouhyoung, M. (2003). On visual similarity based 3D model retrieval Eurographics Computer Graphics Forum.

Johnson, A. E., & Hebert, M. (1999). Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(5), 433-449.

McHenry, K., Kooper, R., & Bajcsy, P. (2009a). Towards a universal, quantifiable, and scalable file format converter. Fifth International Conference on e-Science, e-Science 2009, 140-147.

McHenry, K., Kooper, R., & Bajcsy, P. (2009b). Taking matters into your own hands: Imposing code reusability for universal file format conversion. Paper presented at the The Microsoft e-Science Workshop, Pittsburgh, PA, October 15-19, 2009.

McHenry, K., Kooper, R., Marini, L., & Bajcsy, P. (2010). Designing a Scalable Cross Platform Imposed Code Reuse Framework. Paper presented at the Microsoft eScience Workshop, Berkeley, CA. October 11-13, 2010.

McHenry, K., Kooper, R., Ondrejcek, M., Marini, L., & Bajcsy, P. (2011). A Mosaic of Software. IEEE 7th International Conference on E-Science, e-Science 2011, 279-286.

Ondrejcek, M., McHenry, K., & Bajcsy, P. (2010). The conversion software registry Paper presented at the The Microsoft e-Science Workshop, Berkeley, CA, October 11-13, 2010.

Yeh, T., Chang, T., & Miller, R. (2009). Sikuli: Using gui screenshots for search and automation. Paper presented at the 22nd Annual ACM Symposium on User Interface Software and Technology (UIST), Victoria, BC, Canada, October 4-7, 2009. http://groups.csail.mit.edu/uid/projects/sikuli/sikuli-uist2009.pdf

Preservation of Digital Objects at the Archaeology Data Service

Jenny Mitcham

Archaeology Data Service, University of York, The King's Manor, York, YO1 7EP, UK

Abstract. The Archaeology Data Service (ADS) works to preserve and disseminate the wide variety of data types that archaeologists produce. These range from simple texts, spreadsheets and digital photographs, to far more complex digital objects, such as 3Dimensional digital models and visualisations. Although, for the ADS, the principles of digital archiving remain the same no matter what the file type, complex data, which is often large in size and in proprietary file formats, brings new challenges. This paper will discuss the approach that the ADS takes to digital archiving, with a particular focus on preservation and dissemination strategies for complex objects.

Background to the ADS

The Archaeology Data Service (ADS) was founded in 1996 for the purpose of preserving digital data produced by archaeologists based in the UK, and making it available for scholarly reuse. The ADS was initially established as part of the Arts and Humanities Data Service (AHDS), with sister services covering other disciplines within the arts and humanities. Data are archived to ensure long term preservation, but they are also made available free of charge for download or via online interfaces to encourage reuse[1].

The digital archive at the Archaeology Data Service was established several years prior to the acceptance of the Open Archival Information System (OAIS) model as an ISO standard in 2002 (Lavoie 2004). ADS archival procedures and policies have evolved over time as the organisation itself and the wider world of digital archiving has grown and matured. We have now adopted the OAIS reference model approach and mapped our archival practices to it.

ADS preservation strategy is based on the migration of submitted files into archival formats suitable for preservation. This strategy is detailed in our formal Preservation Policy which, along with all ADS advice and guidance, is available on-line[2]. File formats for preservation are selected using a number of criteria as described in the DPC Technology Watch Report ‘File Formats for Preservation’ (Todd 2009).

It is important for the ADS to be able to demonstrate that it can be trusted as a digital archive. A researcher who has invested time and effort in creating a dataset needs to trust that we will take care of their data in the long term. They also need to trust that they will be able to locate their files whenever they need to access them in the future. In order to further enhance the level of trust that our depositors put in us as an appropriate repository for their data, we successfully applied for the Data Seal of Approval (DSA)[3] at the start of 2011. Undergoing the DSA peer review process has benefited the ADS in many ways (Mitcham and Hardman 2011), but the opportunity it gave us to reflect on and further formalise ADS archival procedures was a valuable outcome in its own right.

[1] http://archaeologydataservice.ac.uk/about/background

[2] http://archaeologydataservice.ac.uk/attach/preservation/PreservationPolicyV1.3.1.pdf

[3] http://www.datasealofapproval.org/

ADS Collections Policy

The ADS has a Collections Policy which describes the types of data that we accept for deposit[1]. Data has to relate in some way to archaeology or the historic environment for it to be of interest to our designated community (as per the OAIS model). Our Collections Policy describes in some detail the types of data that we accept. Regarding visualisations and 3D reconstructions it reads as follows[2]:

> *2.2.8. Visualisation.*
> *3D reconstructions, including computer-generated solid models, VRML, and other visualisations will be collected where it is feasible to maintain them and where they are considered to be capable of reuse and restudy or are seen as being of importance for the history of the discipline in accordance with the procedures defined in the AHDS Guide to Good Practice for Virtual Reality. In general the ADS will also preserve the data from which the model is derived, and sufficient metadata in accordance with the principles of the London Charter (2009).*

Whilst visualisations produced by archaeologists are clearly within our collections remit, it is important to highlight the caveats mentioned in the Collections Policy. Data must be 'capable of reuse' or be 'of importance for the history of the discipline'. These factors can be quite subjective and difficult to quantify. There is mention of 'sufficient metadata' and this is one of the challenges that we frequently face in our day-to-day archiving work. We believe that the single biggest barrier to the future reuse of data is inadequate documentation (Austin & Mitcham 2007). This goes for all types of data that we receive, not just visualisations, but it is often the case that the more complex the dataset, the more metadata is required to successfully archive it.

Despite the variety of data types mentioned in our Collections Policy, much of the data that we ingest into our digital archive is not of a complex nature, the most commonly ingested formats are all widely understood data types with fairly straightforward migration paths to suitable archival file formats. However, there have been a number of projects that we have worked on over the last few years which have brought complex data to the forefront of our minds and enabled us to formulate and put into practice strategies for managing them.

The Big Data Project

In 2005-6 the ADS carried out a project for English Heritage entitled 'Preservation and Management Strategies for Exceptionally Large Data Formats' but more informally known as 'The Big Data Project'[3]. This project aimed to investigate preservation, reuse and dissemination strategies for, what were then considered, exceptionally large data files generated by archaeologists undertaking fieldwork and other research. The project focused on data collection techniques capable of producing large quantities of data during the course of a single survey. These include techniques that result in 3D datasets which are frequently the starting point for the creation of 3D models and visualisations. The

[1] http://archaeologydataservice.ac.uk/advice/collectionsPolicy

[2] http://archaeologydataservice.ac.uk/advice/collectionsPolicy#section-collectionsPolicy-2.2.CollectionDataTypes

[3] http://archaeologydataservice.ac.uk/research/bigData

techniques highlighted by the project case studies were laser scanning, lidar, maritime and terrestrial geophysics, but project recommendations could be applied to all large datasets.

One of the first deliverables of the Big Data Project was the creation of an on-line survey for creators and users of Big Data in archaeology. The survey questions were wide-ranging and covered data creation and reuse, selection and retention policies, file formats, software, archiving strategies and storage. Though the questionnaire was targeted at a relatively small community, the results were interesting and in many cases confirmed existing assumptions. A full report on the findings of the Big Data survey is available on-line[1] but some of the key findings are as follows:

- A wide range of proprietary software packages are in use.
- About half of respondents did not have any policy in place for archiving their data.
- The majority of respondents would be happy to let others access their data.
- The desire for reuse of these types of datasets is considerable.

This project looked specifically at what we termed 'Big Data', but perhaps one of the conclusions that could be drawn from the project is that it is not so much the size of the data that matters, it is the complexity. Though the logistics of moving large files around can be a problem, hardware storage is becoming cheaper over time, computers have increasing processing power, and many archiving tasks can be batch processed or run as background tasks. It is complex digital objects that can be more challenging to a digital archivist than files which are large in size yet relatively simple in structure.

The VENUS Project

The ADS was involved in another project investigating the creation and archiving of large and complex datasets in 2008-9. The European Commission funded VENUS Project[2] aimed to look at methods and tools for the virtual exploration of deep underwater archaeology sites. Underwater sites, normally accessible to just a small number of experienced divers are an excellent use case for 3D modelling where visualisation techniques can help others experience them too. The VENUS project team used underwater recording techniques such as sonar and photogrammetry to collect huge quantities of information about the Roman shipwreck at Port-Miou C off the coast of the Calanques, between Marseilles and Cassis, in the south of France. This information was then used to create visualisations of the site that could be made available to a wider audience. The role of the ADS in this project was firstly to give guidance on how the data should be collected and managed in order to facilitate long term archiving and reuse, and secondly to create an exemplar archive of the data and disseminate the results on-line.

The advice and guidance produced as part of this project is now freely available, both in the form of a VENUS preservation handbook[3] and a Guide to Good Practice for Marine remote sensing and photogrammetry[4].

The exemplar archive published from VENUS project can be viewed on the ADS website[5] and gives an example of how complex datasets can be archived and disseminated in a relatively simple way. Raw multibeam sonar survey data is available to download as

[1] http://archaeologydataservice.ac.uk/research/bigDataQuestionnaire

[2] http://archaeologydataservice.ac.uk/research/venus

[3] http://archaeologydataservice.ac.uk/attach/venus/VENUS_Preservation_Handbook.pdf

[4] http://guides.archaeologydataservice.ac.uk/g2gp/RSMarine_Toc

[5] http://archaeologydataservice.ac.uk/archives/view/venus_eu_2009/

a series of XYZ files. Virtual Reality Modelling Language (VRML) files are available to download in a compressed (zipped) format, and an animation of the VRML model is also provided as a streamed video to allow researchers to preview the model before they commit to downloading a large file. Other supporting data (including photographs, database files and reports) are also presented on-line.

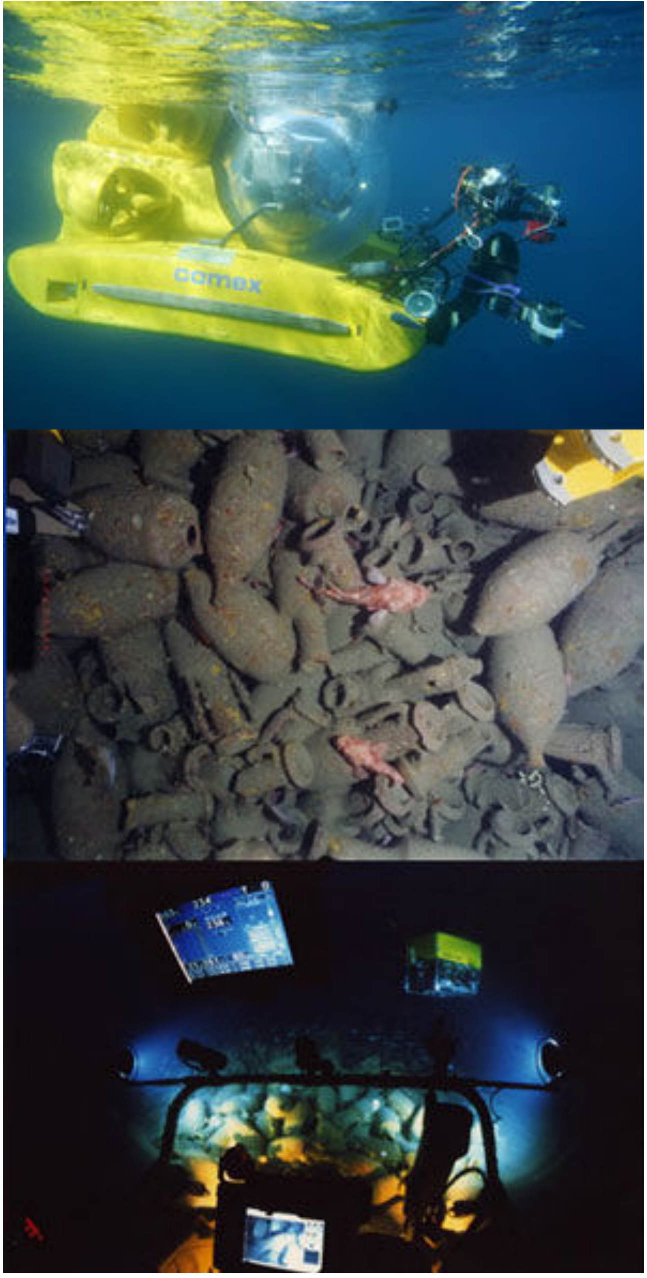


Figure 1. The Remora 2000 during the mission in Marseille, amphora on the Port-Miou C wreck, and the wreck site viewed from the submarine

ADS Archival strategy for complex datasets

The VENUS project exemplar archive as described above is a useful illustration of our strategy for archiving complex datasets. Our primary concern is to archive raw data from which a model is derived. This data should be accompanied by adequate documentation and metadata to ensure its independent utility. OAIS states that an archive should:

> *"Ensure that the preserved information is independently understandable to the user community, in the sense that the information can be understood by users without the assistance of the information producer."*
>
> (Consultative Committee for Space Data Systems 2002)

Raw data and documentation should therefore go hand in hand. There is little value in us preserving a raw dataset from a laser scanning or geophysical survey without the contextual information that makes that data meaningful and suitable for reuse by others.

Our focus on the raw data may sometimes be at odds with the focus of the data depositor, who may consider their 3D model to have more value and interest for others than raw, unprocessed output. We would argue however that it is the raw data which has the greatest long term reuse potential. It may be used for creating alternative interpretations of a site, or be repurposed to answer entirely different research questions long after the platform and software for displaying the visualisation has become obsolete.

Where appropriate the ADS are also keen to archive other types of data produced by a project where a suitable preservation path can be found. These may include processed data, derived data, decimated data, 3D models, visualisations and fly-throughs. We are often inhibited in this field by the large numbers of proprietary and binary file formats used to analyse, process and visualise 3D data, and a stable and open preservation format may not always be available to us[1]. Sometimes these files can only be preserved on a 'best efforts' basis with no guarantee that we would be able to maintain their functionality over time. In these situations the depositor should supply 'snap-shots' of the model (image and movie files for example) alongside full documentation so that future researchers can view the model even if they are not able to explore it first-hand.

When planning selection and retention strategies for data prior to archiving, we find 'Preservation Intervention Points' a useful approach. Preservation Intervention Points or PIPs are a concept that came out of the VENUS project. They represent points across the lifecycle of the data at which the data creator may want to submit data for archiving. Preservation Intervention points will occur at data capture, after processing, post-processing and analysis and at the point of the creation of dissemination outputs. At the end of each of these processes, the data may have been altered in some way to produce a derived dataset worthy of preservation in its own right. Data creation and analysis can be a complex series of processes and this model highlights the fact that preservation copies of the data may need to be created at a variety of different points through the lifecycle of the project.

At each potential Preservation Intervention Point, an assessment must be made to establish whether the data at this stage is worthy of preservation. The criteria for making this assessment are as follows[2]:

- **Preservation Metadata**—There should exist appropriate levels of preservation metadata so that the data are made reusable rather than simply preservable.
- **Resource Discovery Metadata**—There should be appropriate levels of resource discovery metadata so that data from each point can be meaningfully differentiated and distinguished from other parts of the dataset (this mostly applies to legacy data).
- **Identifiable Migration Paths**—There should be clear migration options for data at all stages.

[1] For examples of situations where a suitable migration path cannot be found, see Jeffrey 2010.

[2] These criteria are taken from the Guides to Good Practice: http://guides.archaeologydataservice.ac.uk/g2gp/ArchivalStrat_1-3

- **Reuse Cases**—This is probably both the most important criterion and occasionally the most difficult to judge. Where the data is in a form that can obviously be used by other researchers, or in other contexts, then the question is simply whether reuse is likely to occur. The other complication is that, for certain types of data, a reuse case can be imagined as feasible even if it is currently not being enacted. An example of this would be a form of data that might lend itself to a post-processing technique under development, or merely envisaged as possible in the future (or an enhancement to an existing technique).
- **Repeatability**—Is the process that created this data repeatable? If so, an earlier stage may be an appropriate PIP; if not, then this intervention point should be selected.
- **Retention policy**—The data should match the retention policy of the target archive.
- **Value**—The cost of intervening to preserve data at this particular point, given that no project has an unlimited budget. "Value" here also means the value of the material to be archived, e.g., it might be worth preserving data produced by a repeatable process if that process were particularly expensive and difficult to reproduce. Value, therefore, has to do with balancing the perceived worth of the data against the cost of archiving.

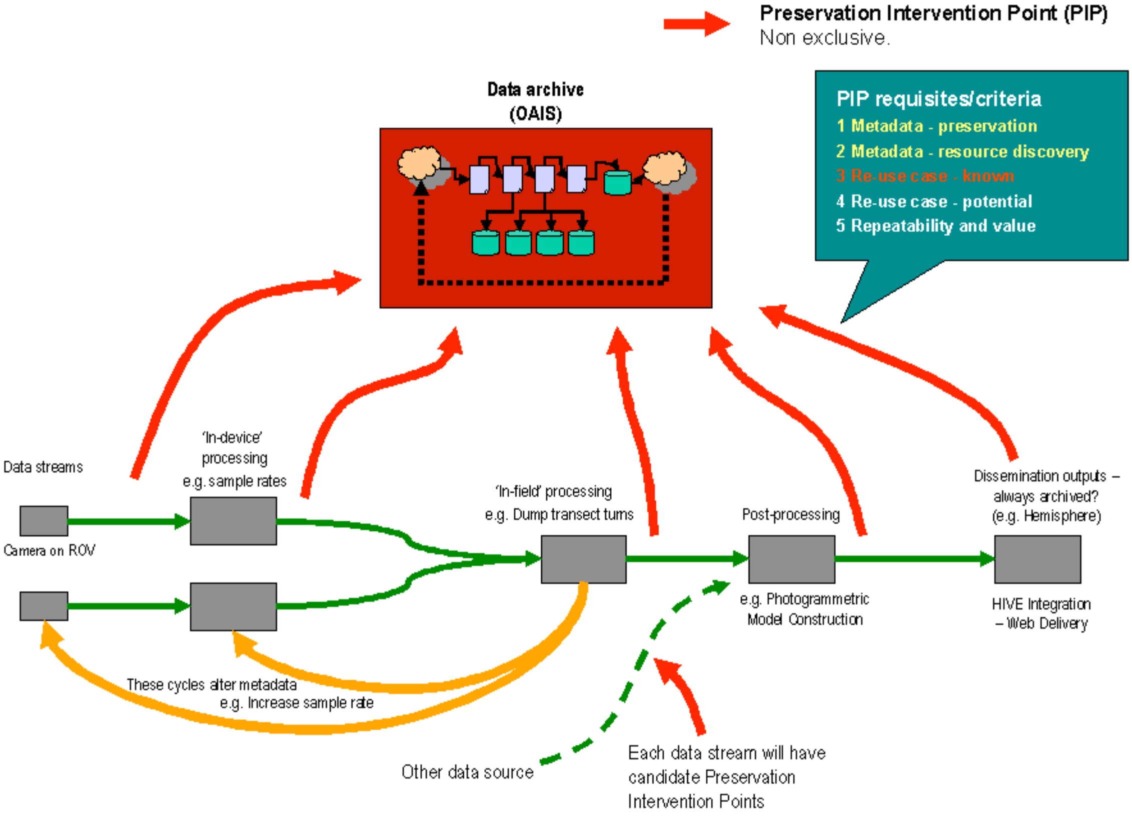

Preservation Intervention Point (PIP)
Non exclusive.
Data archive
(OAIS)
PIP requisites/criteria
1 Metadata - preservation
2 Metadata - resource discovery
3 Re-use case - known
4 Re-use case - potential
5 Repeatability and value
Data streams
Camera on ROV
'In-device' processing e.g. sample rates
'In-field' processing e.g. Dump transect turns
Post-processing
e.g. Photogrammetric Model Construction
Dissemination outputs – always archived? (e.g. Hemisphere)
HIVE Integration – Web Delivery
These cycles alter metadata e.g. Increase sample rate
Other data source
Each data stream will have candidate Preservation Intervention Points

Figure 2. Preservation Intervention Points. Image from "Guides to Good Practice: Data Selection: Preservation Intervention Points, http://guides.archaeologydataservice.ac.uk/g2gp/ArchivalStrat_1-3

Using this approach to work through and formalise selection and retention decision making demands a certain amount of intellectual engagement with the preservation process from the data producer. By encouraging them to think about the 'value' of the data they hold, they can start to collate a subset of files for archiving for which a reuse case can be envisaged.

Guides to Good Practice

The importance of submitting adequate metadata alongside the data to be archived has already been stressed. One of the core functions of the ADS is to offer advice and guidance to data creators and depositors on preparing and documenting a dataset prior to archiving. One of the ways that we achieve this is through our Guides to Good Practice[1]. These Guides were originally published by the ADS and the AHDS from 1998 onwards and were available in hard copy and also free of charge as static on-line publications. Since January 2009 we have been working with both English Heritage[2] and Digital Antiquity[3] to refresh and enhance the Guides to Good Practice series. Through this new, collaborative project we have updated and restructured the original Guides, making them available in an on-line wiki environment to allow easy and quick collaboration and also more frequent future updates. The original Guides included subjects such as excavation, geophysics, Geographic Information Systems (GIS), Computer Aided Design (CAD) and virtual reality but a range of new Guides have now been added to the existing set, including 3D laser scanning, lidar and photogrammetry.

The aim of the Guides is to provide practical advice on the creation, preservation and reuse of digital resources, including valuable guidance on metadata creation. The level of metadata and documentation that should accompany a project archive will vary depending on the nature of the data to be deposited. Metadata for visualisations and simulations should include following[4]:

- Description of the project as a whole (title, dates, funders, copyright, creators etc.).
- Description of the audience and level of interaction (who is the model aimed at and how can they interact with it?)
- Methods and techniques used to create the model (application format, specification, hardware platform, authoring tools etc.).
- Details of datasets that have been incorporated into the model: laser scanning, geophysics, lidar for example. Note that separate guidance on the level of metadata required for each of these techniques will be listed in the appropriate Guide to Good Practice.
- Which features of the model are based on hypothesis rather than evidence.
- Details of delivery platform (OS, web browser, plug-in, hardware, scripting language etc.).
- Description of the look and feel of the model, and what it feels like to experience it.

The last of these points is possibly the hardest to get right. Virtual worlds are notoriously difficult to describe as our experiences of them can be quite subjective – different people may view and explore them in different ways.

[1] http://guides.archaeologydataservice.ac.uk/

[2] http://www.english-heritage.org.uk/

[3] http://www.digitalantiquity.org/

[4] For further guidance on appropriate levels of metadata and documentation see The London Charter (http://www.londoncharter.org/) and the chapter by Hugh Denard in this collection.

Where is all the Complex Data?

At the ADS the framework for the archiving of complex digital datasets is firmly in place. We are a trustworthy digital archive with 15 years of experience behind us. Our research work over recent years includes projects that have specifically looked at large and complex 3D datasets and we have a published set of guidance documents aimed at data creators who are working in this field. However, an analysis of our actual data holdings would show only a small number of datasets that include visualisations, simulations and 3D models. We know that archaeologists are often early adopters of new technologies and techniques for visualising their datasets, but there are a number of reasons why the results of their work are not routinely deposited with us.

At the ADS we charge the data depositor for our services as a digital archive. This is a one off charge that covers the cost of ingest, storage, dissemination, administration and any ongoing file migrations that may be carried out to guard against future obsolescence. Our charging policy is available to view on-line[1] and is based not only on the number of files deposited with us but of their size and complexity. Larger and more complex files are more time consuming to work with and more expensive to store thus will cost more to deposit with us. The benefit of this approach is that people are encouraged to think about selection and retention and consider where the Preservation Intervention Points should be rather than giving us every file that has been created as part of a project. The downside however is that if archiving costs have not been budgeted for, when a project ends, there may be little money left to cover archiving of the full dataset, so a selection of smaller and less complex files such as project reports and images may be prioritised.

As discussed above, adequate metadata is essential if we are to successfully archive complex data. Situations may arise where a researcher wishes to deposit a dataset with us but does not have (and cannot produce retrospectively) suitably detailed metadata. This is particularly true of complex data that may require a more detailed set of technical metadata as well as contextual and resource discovery metadata. As part of the ingest process, digital archivists at the ADS perform various checks on data submitted to ensure that it is independently understandable and suitable for reuse. Where crucial documentation is missing, we may have to discuss with the depositor what elements of their dataset are not suitable for inclusion in the archive.

Another issue that sometimes contributes to the problem is that of copyright and ownership. An archaeological researcher is unlikely to have the equipment and expertise required to carry out an airborne lidar survey of their study area. If they wish to use lidar data to create a 3D model of a landscape, they need to acquire that data from elsewhere. At the end of the project this data cannot then be submitted to the ADS for archiving if copyright is not held by the project team. The owners of the lidar dataset may not accede to our terms and conditions making the data (or datasets derived from it) freely available on-line. An example of this can be seen in the University of Birmingham's North Sea Palaeolandscape Project[2] – a digital archive published by the ADS in 2011. This was an interesting reuse of 3D seismic survey data to generate information on the Mesolithic landscape of the North Sea. The primary dataset that they used to create their 3D model was a mapping of the seabed compiled by Petroleum Geo-Services[3]. This seismic data is commercially very sensitive and for this reason could not be widely disseminated by the

[1] http://archaeologydataservice.ac.uk/advice/chargingPolicy

[2] http://archaeologydataservice.ac.uk/archives/view/nspp_eh_2011/

[3] http://www.pgs.com/

ADS. Similarly, the models the project created based on this data also could not be made publicly available.

Conclusion

The Archaeology Data Service is a digital archive with many years experience of preserving and disseminating the digital data that archaeologists produce. Our OAIS and migration-based archival strategy is openly documented[1] and well-established. The approach that we take to archiving remains the same whether the data is simple or complex. There are however particular challenges that we are faced with when handling complex data.

When archiving 3D models and visualizations, the many different technologies used to create and display them can be a problem. This is a fast evolving field, with many different proprietary file formats in use. Sometimes, there will be no obvious migration pathway that will adequately preserve the significant properties[2] of a model. There is no one-size-fits-all preservation strategy for data of this type and individual projects need to be assessed on a case-by-case basis. These problems however should not be insurmountable. A key element in tackling these issues is to ensure that digital archiving is an essential part of a project plan and that funding for it is built into a project budget from the start. The project team need to start a dialog with the digital archive as early as possible to ensure that a suitable preservation strategy can be agreed and that adequate metadata and documentation is created along the way.

There are many reasons why we feel that visualizations and 3D models should be preserved. Archaeological remains are regularly described as being a finite and non-renewable resource and due to natural and man-made processes some sites and monuments will not be with us forever in their current physical form. Thus a 3D visualisation of a panel of rock art which may be gradually weathering and eroding over time may be the best way we have of experiencing that rock art 100 years from now. Other archaeological remains are accessible only to a small minority (for example cave or underwater sites) and visualizations are the only way that the majority of archaeologists, and the public, will ever be able to experience them. Many more sites we know about as archaeologists are no longer in existence or entirely ruinous and a digital reconstruction that allows people to explore how they might have once appeared is an extremely valuable tool for understanding the past. Preservation of models and visualisations such as these is important to the discipline as a whole and is very much within the remit of the ADS.

References

Austin, T., & Mitcham, J. (2007). Preservation and Management Strategies for Exceptionally Large Data Formats: 'Big Data' Final Report 1.03. 28 September 2007. 47 pp. Retrieved from http://archaeologydataservice.ac.uk/attach/bigData/bigdata_final_report_1.3.pdf

Grace, S., Knight, G., & Montague, L. (2009). Investigating the Significant Properties of Electronic Content over Time: Final Report. 27 pp. Retrieved from http://www.significantproperties.org.uk/inspect-finalreport.pdf.

[1] http://archaeologydataservice.ac.uk/advice/preservation

[2] The InSPECT project defines significant properties as follows: “The characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what they purport to record” (Grace et al 2009)

Lavoie, B. F. (2004). The Open Archival Information System Reference Model: Introductory Guide. Digital Preservation Coalition Technology Watch Series Report 04-01. 20 pp.
Retrieved from http://www.dpconline.org/component/docman/doc_download/91-introduction-to-oais

Mitcham, J., & Hardman, C. (2011). ADS and the Data Seal of Approval – case study for the DCC. Retrieved from http://www.dcc.ac.uk/resources/case-studies/ads-dsa.

OAIS. Reference Model for an Open Archival Information System. Recommendation for Space Data Systems Standard, CCSDS Blue Book. (2002). Retrieved from http://public.ccsds.org/publications/archive/650x0b1.pdf

Stuart, J. (2010). Resource Discovery and Curation of Complex and Interactive Digital Datasets. In C. Bailey & H. Gardiner (Eds.), Revisualizing Visual Culture (pp. 45-60). Farnham: Ashgate.

Todd, M. (2009). File Formats for Preservation. Digital Preservation Coalition Technology Watch Series Report 09-02. 43 pp. Retrieved from http://www.dpconline.org/component/docman/doc_download/375-file-formats-for-preservation.

Case Studies and Discussion Topics

Ecologies of Research and Performance: Preservation Challenges in the London Charter

Hugh Denard

Digital Humanities, King's College London, 26-29 Drury Lane, London WC2B 5RL, UK

Abstract. This paper is a highly edited version of a transcript of a lecture given at the POCOS Symposium at King's College London on 17th June 2011. It outlines the development of the London Charter and delineates its relevance to preserving complex objects, including both digital and non-digital material, particularly 3D visualisations and simulations.

Introduction

I feel somewhat out of my depth speaking about Digital Preservation, being a theatre historian with interests in Greek and Roman antiquity and 20th century Irish drama. Apart from a deep interest in methodological issues concerning three-dimensional visualisation for humanities research (in particular as expressed in the *London Charter*), preservation is a relatively unfamiliar territory for me. But I hope that I can offer something useful by way of problems, challenges and ideas. Those of you who are old hands at digital preservation will be familiar with these kinds of issues, but I hope my contribution, from the perspective of theatre history and digital visualisation, may still be in some way useful.

I would like to begin by discussing the *London Charter for the Computer-based Visualisation of Cultural Heritage*, firstly by asking what the London Charter has to say that is relevant to the preservation of complex digital objects? Secondly, I want to look at the nature of some of the kinds of complex objects (note, not just complex *digital* objects) that are produced by multi-faceted projects – in this instance, projects with both humanities research and artistic dimensions – as well as projects that span digital and non-digital materials, as I think there are huge issues to discuss in these areas.

The London Charter

For those of you who are new to the London Charter, it started off in February 2006 with a symposium and workshop (which, as it happens, included various people here today) on "Making 3D Visual Research Outcomes Transparent" which was held at both the British Academy and King's College London as part of our "Making Space" AHRC ICT Strategy Project. We were discussing the problem that there was, at that time, no international consensus or standard about the type or level of documentation needed to communicate to scholarly audiences the methods and outcomes of visualisation-based research projects. To illustrate: when we create a computer visualisation of an historical monument or object, it is really no more than a pretty picture unless we also communicate to people what has gone into the making of that visualisation: what is the evidence, how reliable is the evidence, what decisions have been made in order to create this digital object we call a visualisation. If we do not provide this kind of information, the visualisation might look compelling as an image in a book or on screen, but it is utterly useless from a research perspective; it cannot be properly understood or evaluated. This is a problem that had been much written about for over a decade before we started putting our minds to it.

At this symposium in 2006, I stood up and said: "listen, we've been writing about this problem in academic articles for years and there doesn't yet seem to be, within the community, a galvanising of effort and energy around actually implementing best practice in this area. So how do we move from this abstract idea of what is good practice to actually creating a *culture* of good practice within the heritage visualisation domain?" I went on to suggest that what we needed was a different kind of publication; rather than yet another article on the subject, we needed something that would call itself a "Charter", which would enable us to draw the community's attention to what would amount to a formal, consensus-based statement of best practice for heritage visualisation. Calling it a Charter would enable us to say: "If you really want your work to be taken seriously by your peers, you really need to live up to the principles that are embodied by this Charter." Everyone agreed this was a good idea, so over the next few days I drafted, with intensive input from an expert working group, the first version of The London Charter, which we released in early March. Then, in February 2009, we published Draft 2.1, which was a much more robust version and which reflects our current thinking on these issues. All the drafts of the Charter, together with its detailed history and an introduction, can be seen on the London Charter website.[1]

The Charter has six Principles, each of which addresses a specific topic. Let me give you a very brief overview of its key recommendations.

Principle 1, "Implementation", addresses the scope of the Charter's applicability – i.e. it identifies when this sort of best practice is relevant. Principle 2 discusses the importance of making aims and methods cohere with each other. Principle 3, "Research Sources", discusses the importance of publishing the evidence upon which research visualisations are based, and of demonstrating that you have systematically and rigorously evaluated that evidence.

As you can already see, it really is not Rocket Science; it is just basic good research practice. If we are making an argument using words, we use footnotes and bibliographies as a matter of course to show the context and provenance of our thinking; we set out our argument, show how it relates to what others have previously said, produce our evidence, and clearly set out our analysis of that evidence. All we are saying in the London Charter is that we also need to transfer that basic good practice to arguments that are made using images, and that setting out our research sources goes a long way towards making our work intellectually transparent.

Then there is Principle 4, "Documentation". This is slightly more difficult because visualisations typically appear either as still images or as models that we can move around. However, the visualisations themselves do not show or explain the processes through which they were created. In a written article, we show, step by step, how our thinking is evolving and how we are using evidence to support our arguments. But if all we publish is an image or a model, it is like publishing only the final paragraph of an article – all conclusions, no argument or evidence. For a visualisation, therefore, we actually have to do a lot of work to tell people how we got there. So it is crucial to document each part of the research process. Drew Baker at King's took the term "paradata" and very usefully applied it to this domain; paradata being information about processes.

[1] http://www.londoncharter.org

There are two final Principles, on "Sustainability" and "Access", which I will discuss in more detail shortly, as they perhaps have the most direct relevance to Digital Preservation as it is conventionally defined.

Before that, though, it is worth noting that the London Charter has had a great deal of impact. It has been translated into a number of languages: Japanese, Italian, French, German, Spanish. The 100-member EU EPOCH Network of Excellence strongly endorsed the Charter, and it has attracted the interest of senior figures within the EU, UNESCO and in a number of national-level bodies. The UK's Archaeology Data Service has set the London Charter as a benchmark for the deposit of digital visualisation; their Collections Policy (4th edition, section 2.2.8, "Visualisation") states that:

> *3-D reconstructions, including computer-generated solid models, VRML, and other visualisations will be collected where it is feasible to maintain them and where they are considered to be capable of reuse and restudy or are seen as being of importance for the history of the discipline in accordance with the procedures defined in the AHDS Guide to Good Practice for Virtual Reality. In general the ADS will also preserve the data from which the model is derived, and sufficient metadata in accordance with the principles of the London Charter (2009).*

This ADS Collections Policy represents an important discipline-based endorsement. The Italian Ministry of Culture has also accepted the London Charter as a guideline for work in their domain. The Charter has also had a lot of impact, and is now widely cited, in scholarly publications. There is also a new set of Principles emerging, called the Seville Principles, which the International Forum of Virtual Archaeology is working on, and which describes itself as an implementation of the London Charter specifically for the domain of archaeology.

Preservation Issues raised by the London Charter

So, let us have a look in a bit more detail at some of these principles as they relate to Preservation issues. The core statement of Principle 4, "Documentation" is:

> *Sufficient information should be documented and disseminated to allow computer-based visualisation methods and outcomes to be understood and evaluated in relation to the contexts and purposes for which they are deployed.*

The phrase "in relation to" is absolutely crucial. The London Charter is not an absolute standard: it is a relational model. If I am doing a quick, five-minute plan using Google SketchUp, for instance, just to work out the potential relationships between two objects in space, nobody is going to require me to supply 50 pages of paradata to go with it. The documentation needs to be proportionate to the visualisation and to the function that the visualisation is fulfilling. By contrast, if I am using digital visualisations to create a formal, academic publication (for example KVL's work on the 100-room Roman Villa at Oplontis which is to be the basis of a born-digital scholarly publication endorsed by the Archaeological Superintendent of Pompeii), then that visualisation has to publish documentation at a very high level of detail. So the Charter places the onus on the creators of the visualisation to work out what quantity and quality of documentation is appropriate according to the community's consensus on what represents best practice.

From a preservation perspective, what we are dealing with is the preservation of, firstly, ephemeral processes, and secondly, the traces that these research processes leave in the form of documentation, which can appear in a wide variety of formats, from video diaries to illustrated scholarly articles.

The twelve ensuing sections of Principle 4 discuss a variety of important ancillary issues, which we can review briefly. These include the importance of taking account of the ways in which the process of creating documentation can enhance research practice (4.1-4.3), and of documenting knowledge claims (4.4) – i.e. making clear whether a particular visualisation is recording evidence or is representing an hypothesis, and if the latter, then establishing how secure that hypothesis is.

The Charter stipulates the need to document research sources, processes and methods (4.5-4.9): to explain why we are using these methods – indeed, why we are using a visualisation-based approach at all: why is three-dimensional digital modelling an appropriate means of addressing our specific research or communication aims? In short, the Charter challenges us to think through, with rigour, and to document our decisions about why we are doing what we are doing, and how we are doing it.

Section 4.10 discusses the documentation of dependency relationships:

> *Computer-based visualisation outcomes should be disseminated in such a way that the nature and importance of significant, hypothetical dependency relationships between elements can be clearly identified by users and the reasoning underlying such hypotheses understood.*

So for example, if all I know is that a glass of water was here in the Anatomy Lecture Theatre at King's College London at 12.30pm on 17th June 2011, in precisely this position in three-dimensional space (expressed as X,Y,Z coordinates), then it follows that there must be some additional object upon which the glass is resting, otherwise the glass would be sitting in mid-air. So the position of the glass depends upon something being present to support it – a dependency relationship. If I am fortunate, my research sources may include a photograph of this lecture theatre showing a fixed desk in this location. If I do not have such a photograph, I might study images of comparable lecture theatres and hypothesize that the glass is likely to be resting on a desk, which may be supported by a review of types of objects which are both likely to be in this position and which would also result in a surface of the appropriate height. Put simply, each piece of evidence may imply certain other things, but in order for our visualisation-based arguments to be intellectually transparent, we need to document and to publish those detailed chains of reasoning.

The sections of the Charter covering the documentation of formats and standards (4.11-4.12) also relate to preservation issues, stipulating that:

> *Documentation should be disseminated using the most effective available media, including graphical, textual, video, audio, numerical or combinations of the above. (4.11)*

and:

> *Documentation should be disseminated sustainably with reference to relevant standards and ontologies according to best practice in relevant communities of practice and in such a way that facilitates its inclusion in relevant citation indexes. (4.12)*

The Charter aims to provide enduring, methodological principles, and so deliberately avoids entanglement in technical details which are likely to become obsolete over time. So, without dwelling on the above two articles, we may note that they remind us of the importance of developing discipline-specific implementation guidelines.

There is a main theme in today's presentation that I want to highlight, which is that documentation is a means of preserving the ephemeral, and often intangible, aspects of highly complex research processes. These three-dimensional digital models are themselves complex digital objects which have technically challenging digital preservation requirements. But equally, documenting in detail the process through which we produced them is also an important aspect of preservation – we need to create surrogates, through documentation, for research processes that unfolded over time. Of course, it is of central importance to determine what we think is important to preserve, which allows us to work towards a preservation strategy, and I think the London Charter provides a very strong framework for identifying priorities for documentation/preservation.

Let us move on now to look at Principle 5, "Sustainability", which opens as follows:

> *Strategies should be planned and implemented to ensure the long-term sustainability of cultural heritage-related computer-based visualisation outcomes and documentation, in order to avoid loss of this growing part of human intellectual, social, economic and cultural heritage.*

Here are some key concepts:

> *The most reliable and sustainable available form of archiving computer-based visualisation outcomes, whether analogue or digital, should be identified and implemented. (5.1)*

If we are concerned that digitally preserving our visualisations may prove difficult – whether for technical reasons or through lack of infrastructure, resources or active data management – if we think that our digital data may be under threat in the future, then we need to think about creating some sort of physical print, whether as an actual three-dimensional object through 3D printing or as a set of print images. If we cannot guarantee that we will be able to preserve the actual digital data, we can at least preserve – and *should* preserve – an enduring, physical record of our visualisation.

This is perhaps not the message we want to hear in the present symposium, because we are concerned with the digital preservation of the data and content itself. But I think it is an important point to put on the map, because not everyone has access to the resources or the expertise to carry out these complex preservation tasks, and not all institutions have the resources indefinitely to maintain them. So while we should keep pressure up, at policy level, to develop and advance the digital preservation agenda, we also need to think carefully about what other surrogates, including hard copy surrogates, we can create as a fall-back strategy.

> *Digital preservation strategies should aim to preserve the computer-based visualisation data, rather than the medium on which they were originally stored, and also information sufficient to enable their use in the future, for example through migration to different formats or software emulation. (5.2)*

This recommendation was really designed for people within the humanities and heritage domain who had not yet conceptually made the distinction between medium and content; the aim was to get them to think about whether it really mattered if a CD, for example, would still be usable in 10 years' time, or whether the important question was whether the bits would still be available and usable.

> *Where digital archiving is not the most reliable means of ensuring the long-term survival of a computer-based visualisation outcome, a partial two-dimensional record of a computer-based visualisation output, evoking as far as possible the scope and properties of the original output, should be preferred to the absence of a record. (5.3)*

Or, to paraphrase, if we have to choose between no preservation at all and a hard copy preservation, please can we have the hard copy.

> *Documentation strategies should be designed to be sustainable in relation to available resources and prevailing working practices. (5.4)*

This is, as much as anything, a nudge to the institutions to say we need those resources and working practices that enable us to carry out this best practice.

Preservation of Mixed Reality Objects

The Charter's distinction between digital preservation and the preservation of digital objects (or between medium and content) is important, but it also brings us to the threshold of an even knottier set of problems, which is my second major theme, namely: what do we do when the object of preservation itself may have non-digital, as well as digital elements? Some of our "complex objects" live in both physical and digital worlds. In such cases, how do we even define the object of preservation?

To give an example: if you go downstairs to the Great Hall today, you will see an artwork that Michael Magruder and I created a year ago called "Vanishing Point(s)" (Figure 1).

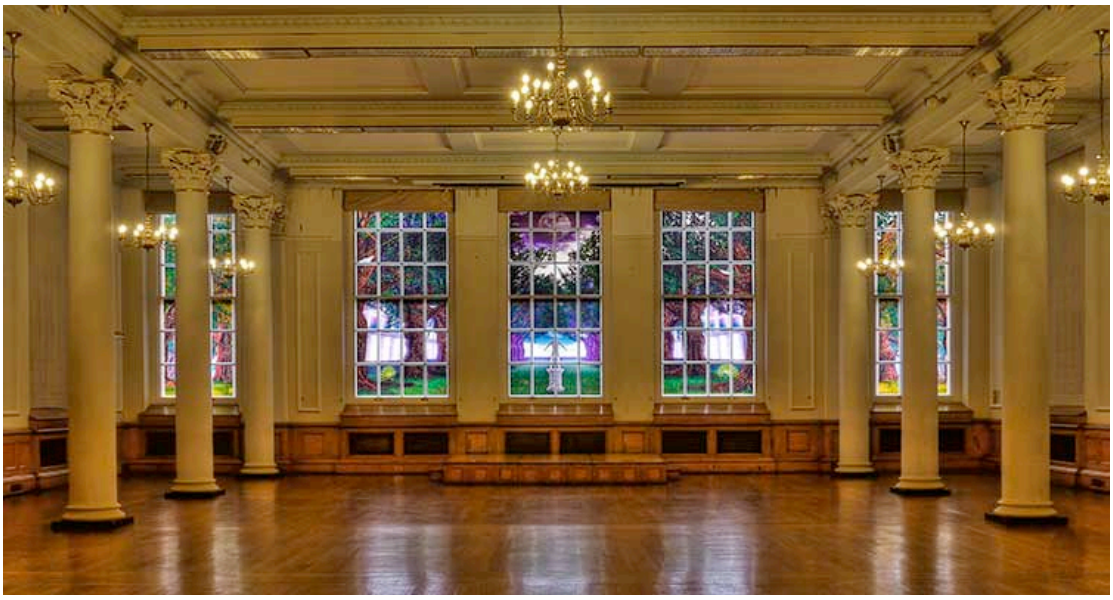


Figure 1. Vanishing Points(s) An artwork by Michael Takeo Magruder and Hugh Denard.
The Great Hall, King's Building, King's College London, July 2010

This artwork[1] is based on a large-scale Roman fresco (Figure 1) dating from the middle of the 1st century BCE, which is to be found *in situ* within the Villa at Oplontis (modern day Torre Annunziata, near Pompeii). This beautiful and complex fresco offers an imagined vista onto a sacred precinct, enclosed by monumental architectural colonnades and, in the foreground, a complex set of architectural framing features – including grand, golden columns, screen walls, coffered ceilings and arches – dressed with significant objects, such as theatre masks, peacocks, burnished shields, paintings. The entire composition is bookended by a quiet garden colonnade which subtly bends the apparent spatial coherence of the whole.

[1] Vanishing Point(s) by Michael Takeo Magruder & Hugh Denard. http://www.takeo.org/nspace/sl005/. The next volume from this series (The Preservation of Complex Objects. Vol. 2. Software Art) presents in depth issues related to the preservation of digital artworks.

Vanishing Point(s) translates the compositional principles and the spatial rhythms of this ancient fresco into a different kind of space – that of the Great Hall at King's. Taking the idea that the Roman fresco invites the viewer to enter into a kind of imagined, virtual world, the contemporary artwork is created by taking a snapshot of a fully realised three-dimensional environment specially created within the Virtual World of Second Life; at the time the piece was created, one could go online, as an avatar, walk around this virtual park in real time. This snapshot of the virtual realm was printed out onto 30m^2 of large-format digital transparency film, called Duratrans and applied to one-hundred-and-eight panes of the five-window array of the Great Hall's east wall, achieving a stained glass-like effect. Once installed, the trees in the artwork carried forward into virtual space the rhythm of the actual columns of the Great Hall, overlaying its architecturally constructed garden vista upon the actual view, which is of a drab, congested urban courtyard. In several subtle ways, which I will not detail here, the new artwork draws upon the compositional principles of the ancient Roman fresco, involving an intricate network of social, philosophical, optical as well as aesthetic considerations.

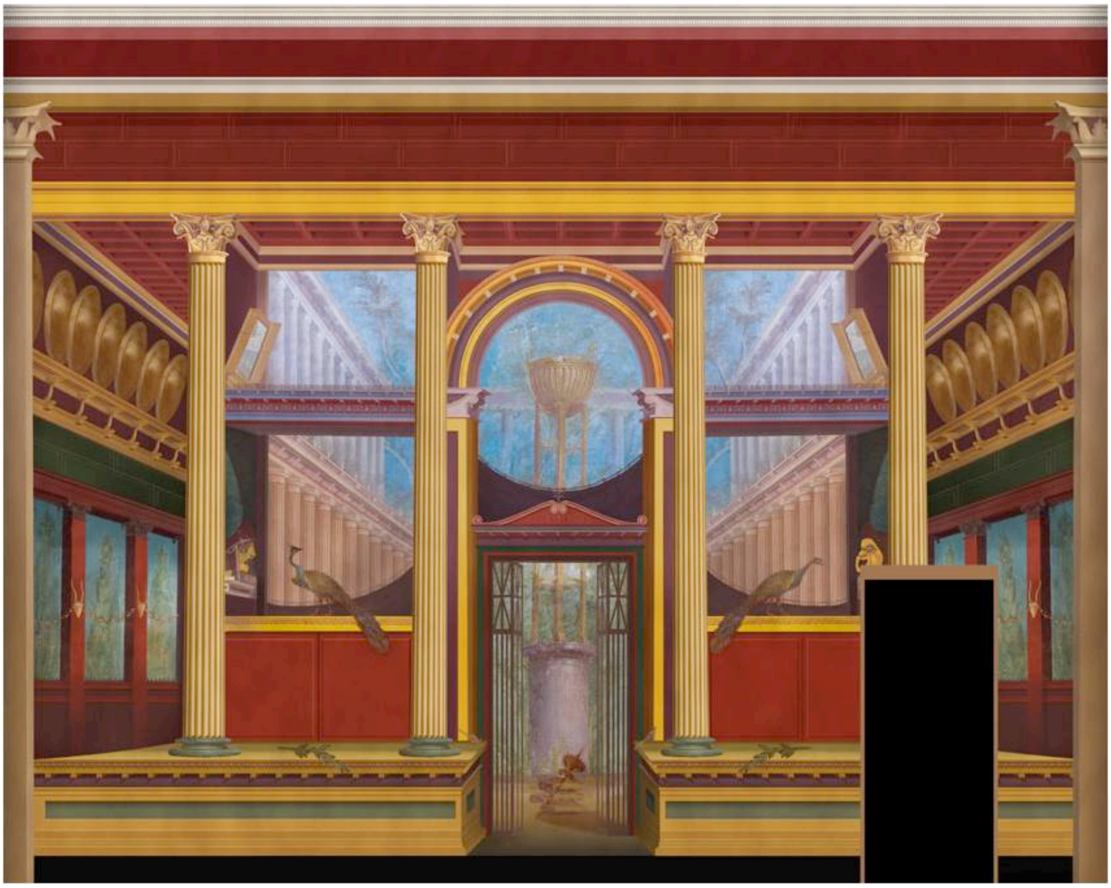


Figure 2. Digital restoration by Martin Blazeby of fresco from Room 15, Roman Villa at Oplontis.
© King's College London 2011.

This particular artwork constructs a crossing point between humanities research (on Roman fresco art) and artistic practice (specifically, Michael Takeo Magruder's work combining elements of physical installation and virtual worlds technologies), and this intersection raises all kinds of issues regarding how we scope and prioritise our preservation processes. How, when we are dealing with a "mixed reality" artwork such as this which comprises a blend of actual and virtual elements (a virtual landscape in Second

Life and a physical installation in the Great Hall), do we even begin to define what is the object that is to be preserved? And who do we identify as the intended beneficiaries of preservation, which would enable us to establish appropriate preservation priorities? How does London Charter-articulated best practice apply to blended entities such as these?

We might consider whether the question of intellectual transparency ought to have any purchase on our preservation strategy. The Preamble to the Charter is very clear that the Charter:

> *…is concerned with the research and dissemination of cultural heritage across academic, educational, curatorial and commercial domains. It has relevance, therefore, for those aspects of the entertainment industry involving the reconstruction or evocation of cultural heritage, but not for the use of computer-based visualisation in, for example, contemporary art, fashion, or design. (Preamble)*

However, Vanishing Point(s), while it provides a pleasant, artificial view for the Great Hall, also becomes much more richly meaningful if its audience is aware of its intense, intertextual relationship with the Roman fresco from the Villa at Oplontis – a relationship, however, which the artwork itself does not overtly present to its viewers. And I hope I have indicated how an artwork such as this becomes even more stimulating and enjoyable if you, the viewer, has access to some of the thinking that explains how the new artwork draws on a deep, multi-faceted interpretation of that Roman fresco and its social, cultural, as well as 'spatial' presence. And there are yet further strands to do with the reception of the artwork, such as the decision by the College authorities to extend its initial 6-week life indefinitely, all of which is part of the "object" which merits consideration as part of a preservation strategy.

So perhaps what this suggests is that preservation has a role, not only in recording, but also in creating what almost amounts to a new iteration of the artwork in the 'medium' of documentation: documentation which not only records, but also actually augments the artwork in a significant way.

Preservation, Access and Enhancing Practice

Let us go on, now, to the final principle of the London Charter, which deals with issues of "Access". I have to admit, I just made up this principle, back in 2006, because I thought it ought to be there. All of the other principles of the London Charter came after consultation with the wider group; but this one, I just drafted and circulated, and was glad to find that others also felt that it ought to be endorsed. The headline principle on Access reads as follows:

> *The creation and dissemination of computer-based visualisation should be planned in such a way as to ensure the maximum possible benefits are achieved for the study, understanding, interpretation, preservation and management of cultural heritage.*

I think this chimes very nicely with what William Kilbride was talking about in his workshops yesterday: that the point of preservation is not to be able to pat ourselves on the back for having avoided risks, but rather we are working in this area because there are opportunities – there is value to be added, and there are all sorts of new, intrinsically worthwhile things we can achieve through preservation – and this, I think, is what the Charter's principle on Access is pointing out as well.

All of which brings me to my third major theme, which is: how can preservation (including, of course, preservation of ephemeral processes through documentation) contribute additional, highly valuable dimensions to historical research and creative practice?

Let us review the sub-sections of the Access principle:

> *The aims, methods and dissemination plans of computer-based visualisation should reflect consideration of how such work can enhance access to cultural heritage that is otherwise inaccessible due to health and safety, disability, economic, political, or environmental reasons, or because the object of the visualisation is lost, endangered, dispersed, or has been destroyed, restored or reconstructed. (6.1)*
>
> *Projects should take cognizance of the types and degrees of access that computer-based visualisation can uniquely provide to cultural heritage stakeholders, including the study of change over time, magnification, modification, manipulation of virtual objects, embedding of datasets, and instantaneous global distribution. (6.2)*

Most cultural heritage visualisations are carried out by publicly-funded bodies; by researchers and museums, and so on, which secure government grants to carry out this kind of work. It seems to me that we therefore have a moral obligation to make sure that we share the benefits of that work with the public. We have an obligation to think about who will benefit from our work, whether school children, tourists, museums or the general public.

I am going to conclude by presenting a case study of a visualisation of the Abbey Theatre in Dublin. The Abbey Theatre, founded by William Butler Yeats, Lady Gregory, *AE*, Edward Martyn, John Millington Synge, opened on 27th December 1904, in a theatre leased and renovated for the purpose by Annie Horniman, which remained in use until fire damaged in 1951. The Abbey (which today occupies a new, purpose-built theatre on the same site) went on to become a world famous theatre, producing playwrights and plays of major international importance.

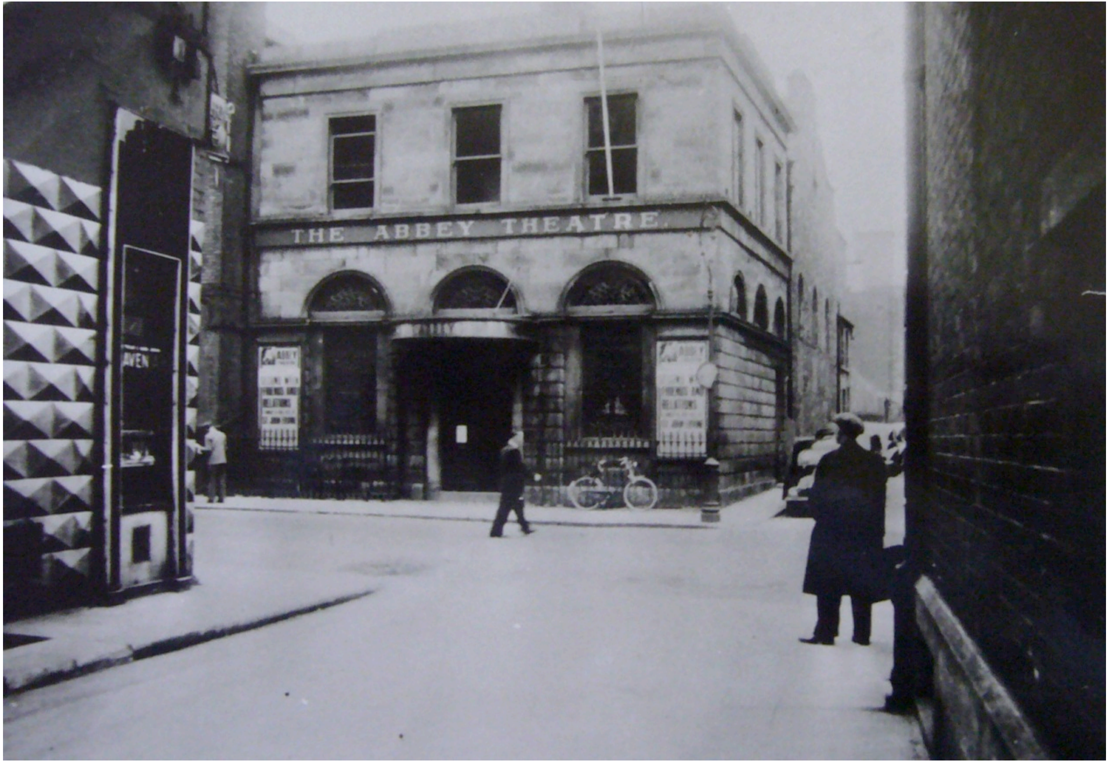


Figure 3. The old Abbey Theatre, Marlborough Street, Dublin. Source: Irish Architectural Archive.

Everyone who goes to school in Ireland studies these plays; I recall studying Sean O'Casey's *The Plough and the Stars* and *Juno and the Paycock*, and J.M. Synge's *The Playboy of the Western World* and *Riders to the Sea*, to name but four. These are major landmarks of international drama, and yet in all of the years that I studied these plays, both at school and subsequently as a theatre student at university, no one once ever asked what now seem to be pretty fundamental questions, such as: what was the theatre like, for which these plays were written? Is there any relationship, perhaps, between the tiny size of the original Abbey's stage and the intimacy of the scenes that the early Abbey playwrights created, including in the new genre of the "peasant play"? Historians have tended to be preoccupied by the "Big Questions" of the Abbey's contribution to an emergent National (Irish) Identity at the beginning of the 20th century, as well as considerations of the literary, poetic and dramatic merits of the plays written for the theatre. These are, in a sense, just variations on the same questions that Yeats himself established as worthy of attention, such as in his poem, Man and the Echo, which, referring to the billing of his play *Cathleen Ni Houlihan* in the Abbey in the week of the Easter Rising, asks: "Did that play of mine send out / Certain men the English shot?" Scholars have found it very difficult to break sufficiently free of the preoccupations proposed by Yeats to investigate these events and materials from significantly different perspectives.

One of these relatively neglected perspectives is space. From January to April 2011, while a Visiting Research Fellow in Trinity College Dublin's Long Room Hub, I investigated the physical properties of the original Abbey Theatre. Working with a local digital communications company, Noho, I undertook to collect and synthesize into a three-dimensional digital model all the evidence I could find about the earliest phase of the theatre's stage and auditorium.

The research sources contributing to the visualisation effort were quite diverse including: a painting by Jack B. Yeats (brother of W.B.) of the pre-renovation theatre; the original, highly detailed, architectural plans for the redesigned theatre by Joseph Holloway (who was also notable for having recorded, in 221 heavy volumes, his extensive observations on the cultural and theatrical scene of Dublin in the first half of the twentieth century); black-and-white photographs of the interior of the early Abbey in the present-day Abbey Theatre's archives, as well as original portraits and artefacts that survived the fire of 1951, published descriptions of the theatre from varying dates; previously unpublished fire insurance maps showing the dramatic changes that the site underwent in phases between 1893 and 1961, with the theatre company's holdings expanding to include a variety of neighbouring buildings used as dressing rooms, green room, properties and scenery stores and so forth. These maps also clearly indicate the social stratification that the new theatre design introduced; in place of a single entrance for the audience of the previous, popular, so-called Mechanics' Theatre, Abbey patrons with tickets to the Stalls and Balcony now entered through an attractive Vestibule on Marlborough Street, while the cheap seats, known as 'The Pit', were accessed from an entrance on Abbey Street, passing through narrow corridors to enter at the very rear of the theatre.

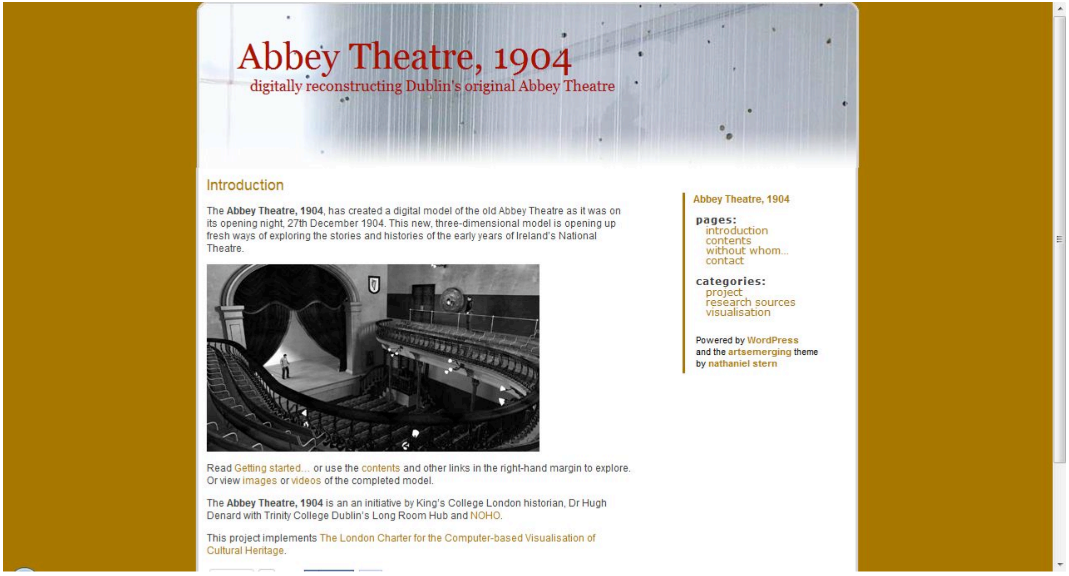
The Irish Architectural Archive on Merrion Square contains drawings from a survey of the theatre by Michael Scott Architects carried out in 1935, which gives additional details that Holloway's plans omit, which are particularly useful when, as often occurs, it is difficult to ascertain which one of the many designs that Holloway created were actually implemented in the theatre. The original programmes from the Abbey, of which the Trinity College Library has a good collection, give us invaluable information about the tradesmen and companies that carried out the electrical installation, scene painting, upholstery, and so on. Honora Faul, Curator of Prints and Drawings in the National Library of Ireland, discovered in Joseph Holloway's vast collection of theatrical ephemera ticket envelopes from the early Abbey, which contained miniature plans of the seating in the theatre, showing the position and number of each individual row and seat, which, although they appear to contain some discrepancies with the early photographs – indicating that they are from a slightly later phase in the building's life – represent yet another extraordinarily valuable source of information.

My research into the history of the fabric of the original Abbey Theatre was full of these unexpected windfalls, of which, none is more extraordinary than the following story. In 1961, ten years after the theatre building had been abandoned due to fire damage, it was finally decided to clear the site to make way for a new purpose-built theatre. At the time, former Dublin City architect, Daithi P. Hanly, realised that no steps were being taken to preserve for posterity the original, historic theatre building, despite its unquestionably iconic importance. He therefore persuaded Christy Cooney, the demolition contractor, to number the external stones and to remove them to Hanly's own garden, in Killiney, south county Dublin, where he would hold them in trust until a way could be found to restore the theatre's façade. Despite a strenuous campaign of several years, backed by numerous luminaries of the Irish theatrical world, the stones remain, to this day, in the late Daithi P. Hanly's garden, where his family continues to care for them in the hope that someday his great vision will be realised. When I visited the garden, and spoke to his family, a member of the film crew accompanying me tilted back one of the stones to find the original lettering of the Abbey Theatre still clearly visible, showing not only the green and cream livery of the later Abbey, but also the gold and red which, subsequent research verified, were the original theatre's decorative colours. In outhouses within the grounds of Hanly house, we found stored the original wooden window-frames, doors, and even billboards that flanked the Vestibule doors. Even the cash till for the Peacock Theatre – the smaller stage – has survived, thanks to Hanly's intervention. And there, on a side table, we found the detailed, scale wooden model of the original Abbey, thought to be lost, which had been photographed and published by James W. Flannery in 1979 in his book, *W. B. Yeats and the Idea of a Theatre*; again, a wonderfully rich research resource.

So, as you can imagine, a project such as this, with such diverse research sources, throws up quite a variety of documentation and preservation challenges. Wishing to ensure that the project represented an exemplary implementation of the London Charter, I needed to devise an approach to achieve intellectual transparency. The key to intellectual transparency is to capture the research process, rather than just the research outcomes, so it was important to maintain some kind of running record of the research and modelling activities. But factoring in, also, the Charter's recommendations on Access, it was equally important that this documentation should be publicly visible. The most logical solution was to create a project blog, which I did.[1] From day to day as the research process unfolded, I was able to write up a fairly informal account of the process, showing readers, step by step, how the research process fed into the digital model that my collaborators, Noho, were creating. I also created a Twitter account for the project, which I used to spread the word about the project and to notify people when the blog had a new update. I felt like the project had really "arrived" when Fiach MacConghail, the Artistic Director of

[1] http://blog.oldabbeytheatre.net

the present-day Abbey Theatre quite spontaneously started re-tweeting my tweets about the old Abbey project!



Abbey Theatre, 1904

digitally reconstructing Dublin's original Abbey Theatre

Introduction

The **Abbey Theatre, 1904**, has created a digital model of the old Abbey Theatre as it was on its opening night, 27th December 1904. This new, three-dimensional model is opening up fresh ways of exploring the stories and histories of the early years of Ireland's National Theatre.

Read Getting started... or use the contents and other links in the right-hand margin to explore. Or view images or videos of the completed model.

The **Abbey Theatre, 1904** is an an initiative by King's College London historian, Dr Hugh Denard with Trinity College Dublin's Long Room Hub and NOHO.

This project implements The London Charter for the Computer-based Visualisation of Cultural Heritage.

Abbey Theatre, 1904

pages:
introduction
contents
without whom...
contact

categories:
project
research sources
visualisation

Powered by WordPress
and the artsemerging theme
by nathaniel stern

Figure 4. The Abbey Theatre, 1904 Project Blog. Screenshot: Jan 2012.

Back to the blog; I created a simple title page giving a sense of what the project is about, and what the blog contains, and foregrounding London Charter compliancy, but saving a fuller description of the background to the project and to the original Abbey Theatre for a separate page for those who wished to drill down a little. Entries are categorised either “Project”, “Research Sources” or “Visualisation”, but can also be found in a manually-created chronological index of entries, with journalistic-style bye-lines giving a sense of their content. Under the “Visualisation” category appear the entries created by the project’s 3D modeller, and founder of Noho, Niall O hOisin, which are brief video diaries in which he describes each day’s work and how information extrapolated from the various research sources were incorporated into the geometry of the digital model; a classic implementation of London Charter documentation / paradata.

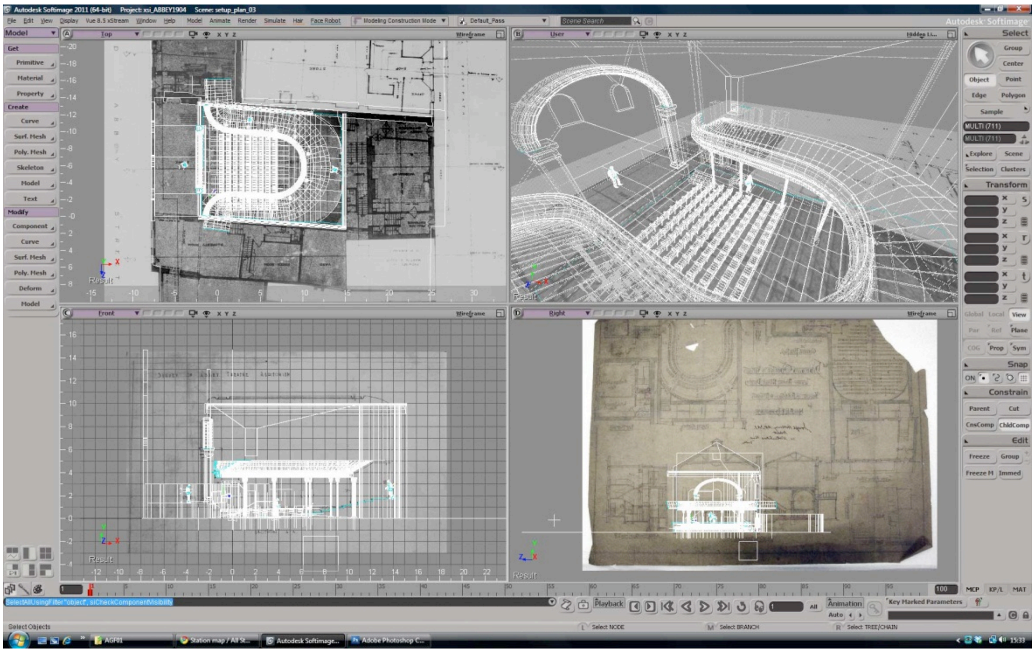

Autodesk Softimage 2011 (64-bit) Project: xsi_ABBEY1904 Scene: setup_plan_03
Top
User
Front
Right
Select
Transform
Snap
Constrain
Edit

Figure 5. Screenshot of the 3D modelling environment incorporating primary sources.

There is also an entry about the project methodology, explaining the approach I was taking in some detail. This included the aim to create three different versions of the model, a Forensic Massing Model, an Artist's Impression and a Forensic Textured Model. The massing model would be un-textured but, colour-coded green, amber and red – a "traffic-light" system proposed by Daniël Pletinckx – to show the different levels of probability ("certain", "probable" and "plausible") accruing to different parts of the model. The Artist's Impression model was to give an intuitive sense of how the theatre may originally have appeared, while the Forensic Textured model was to indicate levels of probability about decorative, rather than architectural, elements. As it happens, we did not in the end realise this final version, because, much to our surprise, we were unable to find a single colour photograph of the interior of the original auditorium, despite its having survived until 1951 – well into the age of colour photography. Consequently, we modelled and rendered the theatre in greyscale, in the style of a black-and-white photograph. (It was several months after the modelling process was over before I happened upon my first textual description of the original colour scheme of the auditorium.) In keeping with this spirit of Access, towards the end of the project, I published, through the blog, a freely downloadable Powerpoint presentation with these rendered images of the model to make it easy for others to incorporate the project's outcomes into their own historical and cultural narratives.

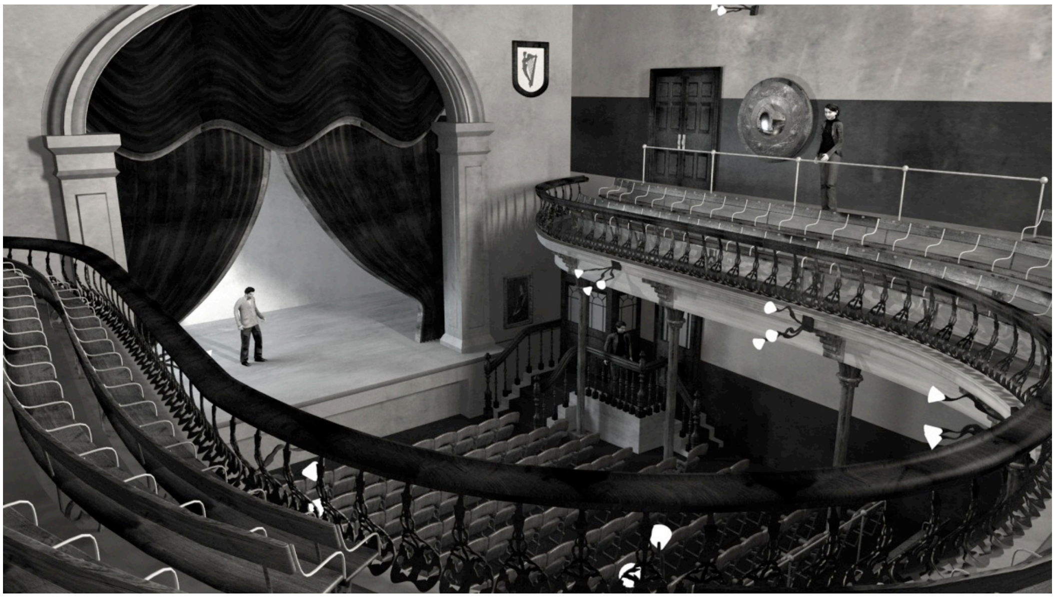


Figure 6. View from the balcony of the digital model of the old Abbey Theatre.

As it turned out, a blog was also a great way of engaging and rewarding stakeholders; I needed lots of people to come on board and help me, from the National Library, the Irish Architectural Archive, Trinity College's Library and Map Library and, not least, the archive at the Abbey Theatre itself. If the project was to be a success, I needed all of these places to collaborate and make things happen in a very short time-frame. So when I had a meeting or a visit that turned up something interesting or exciting, I would blog about it, and would forward the link to my contact at the institution involved, and very quickly I found that they themselves were tweeting the blog to all their friends and colleagues, delighted to be getting such great publicity out of their contribution to the project.

So very quickly, a social ecology began to circulate around the project, with people getting very enthusiastic and motivated to help, as well as established scholars quickly recognizing that this was a new way at looking at this period of Irish cultural and theatrical history; one which highlighted the importance of thinking seriously about the architectural, material and spatial aspects of historical and literary/dramatic phenomena. And because of the old Abbey's key symbolic place within narratives of national identity and nationhood, the project was even taken up by the media, with an illustrated feature on the project appearing in the *Irish Times*, a radio interview with me and Niall on RTE's Lyric FM, as well as additional articles in an online arts magazine, (Christopher Collins, "SHIFTS in Sedition" *Vulgo*, February 2011), and even in a tourist industry magazine, *Ireland of the Welcomes* (Summer 2011). In fact, that the latter was part of a two-page spread of illustrated stories deemed of "national interest", sharing billing with the historic visits to Ireland of President Obama and Queen Elizabeth II says something about the status of the Abbey within the Irish cultural imagination. Together, these indications of both professional and wider public engagement provide a striking example of how documenting and publishing research and visualisation processes can add very significant value, and benefit, to a project, helping it both to gain a lot of momentum in a very short amount of time, and to reach a wide range of stakeholders far beyond those one might have imagined at the start of the project.

But while there are clearly a set of interesting preservation issues around a blog such as this – with its combination of hypertext, still and moving images, drawn from a wide range of sources, and using the proprietary WordPress technology – to my mind, the more intriguing challenge is that of capturing and preserving the social ecology of the research process; that complex set of relationships and interactions that circulated around the project. These, too, are highly important parts of this object called "The Abbey Theatre, 1904 Project".

To illustrate just how significant this challenge is, let me sketch for you one final "moment" in the project. On 15 April 2011, we held an open-air, official project launch at the Samuel Beckett Centre, attended by around 200 people, including those who had contributed funding, as well as time, resources and expertise, as well as members of the wider academic community and of relevant cultural institutions. While guests perused the commemorative postcards and programmes created especially for the event, Prof. Steve Wilmer, Head of the School of Drama, Film and Music, and then Provost of Trinity, Dr John Hegarty, spoke about the project, putting it into the context of Trinity's flagship Creative Arts Technologies and Culture initiative, after which I introduced the project in more detail, with an outdoor data projection playing an animation of the model in the background.

These formal presentations were a prelude to an artistic intervention. Along with my research for the visualisation project, I had been coordinating the creation of a new theatrical performance, designed – along the same conceptual lines as the Vanishing Point(s) artwork discussed above – as a creative companion to the historical research process. Directed by Dan Bergin, *SHIFT* was a devised response, in performance, live music, video mixing, sound and light engineering, to the stimulus provided by both the new digital model and the notorious riots that took place within the old Abbey Theatre, in January 1907, in response to the first performances of J. M. Synge's iconoclastic play, *The Playboy of the Western World*. Lasting just under an hour, it incorporated vignettes of Abbey architect and diarist, Joseph Holloway, period cinema segments, a new musical score by members of silent cinema band, 3epkano, audience participation, and intertextual allusions to Synge's masterpiece. Without going into any detail, you can well imagine that there are as many stories to be told, and captured, about this theatrical facet of the project as there are about the visualisation-based research process. We were fortunate that, in addition to my own blogging, documentary maker Colin Murphy had become interested in the project, and for some weeks had been tracking the progress of the project in both its research and performance aspects; we hope in time to find the resources to edit these up into a form suitable for publishing as short videos on the project website.

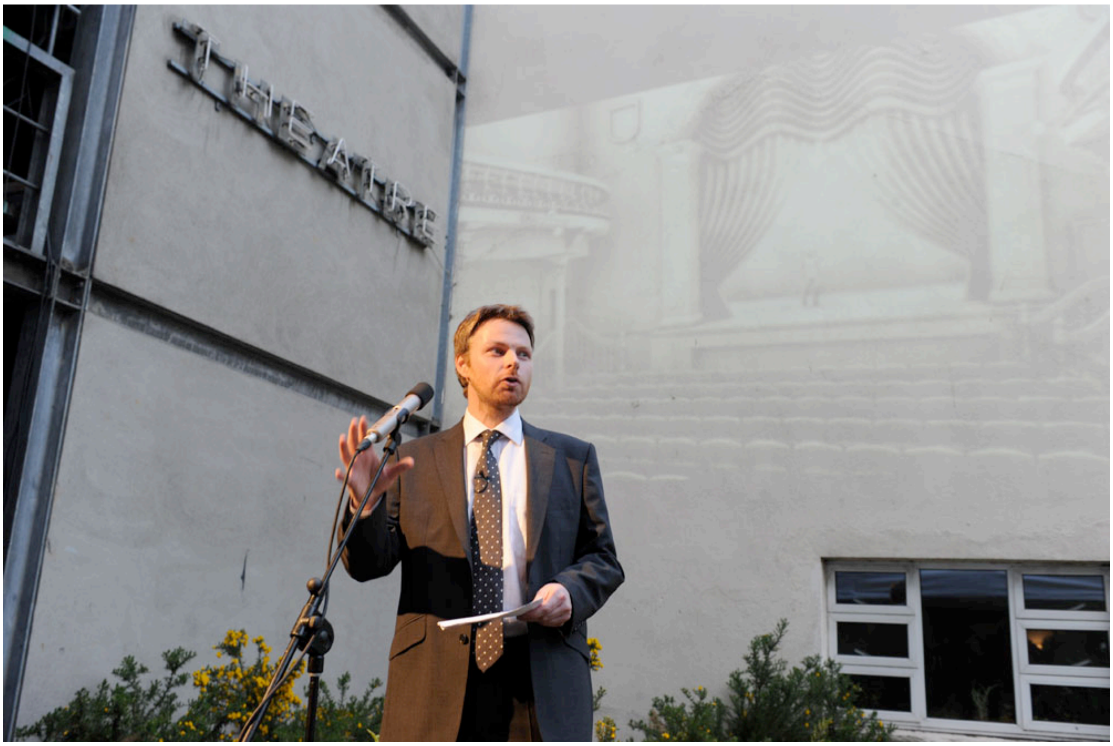


Figure 7. Denard with outdoor projection of old Abbey Theatre animation by Noho. Project launch, Samuel Beckett Theatre, Trinity College Dublin, April 2011. Photo: Sharppix, Dublin.

So this composite launch event/performance, too, and its constituent elements – publicity, scripts, projections, catering, ticketing, performance, video footage and so on – now become part of the social and material ecology of the project. In however small a way, it matters, I think, who was there and what was said. It matters, for example, that Daithi P. Hanly's grandsons, who must have played on those stones in the garden when they were children, were present; it matters, I think, that they were seeing that childhood playground transform itself into elements within a new narrative of the National Theatre. And it matters, perhaps, that people from the Abbey Theatre itself were there, including the Chair of the Abbey's Board, along with representatives of the various libraries and archives who had contributed to the project, and noted scholars and journalists who had previously written and broadcast on the history of the early Abbey. These are energies that we can pour into new initiatives, for example to create a new set of research and digitisation initiatives, or a new generation of educational resources, or to attempt to revive Hanly's vision of the physical restoration of the original theatre façade, or further iterations of the *SHIFT* production concept. But they also matter, quite simply, in their own right, as part of the story to be told.

Such things are all highly intangible and ephemeral, but they are also essential, important parts and outcomes of the object called "The Abbey Theatre, 1904 Project". So what is the role for preservation in a case such as this? I'd simply like to conclude by suggesting that this is one of the challenges to which we could profitably turn our attention, exploring how we might define and undertake the preservation of "objects" that involve a combination of digital and non-digital, tangible as well as intangible, artistic as well as scholarly facets. Processes and relationships need to be captured, or evoked, just as much as outputs. In short, should we ask what might be entailed in preserving the priceless, but elusive, social ecologies of our processes of exploration?

References

London Charter (2009) The London Charter for the Computer-based Visualisation of Cultural Heritage. V. 2.1 (February 2009). Retrieved from http://www.londoncharter.org/

Preservation of Complex Cultural Heritage Objects – a Practical Implementation

Daniël Pletinckx

Visual Dimension bvba, *Ename,B-9700, Belgium*

Abstract. 3D visualisations of complex cultural heritage objects, such as archaeological objects, historical landscapes, buildings and other man-made structures of the past, are more and more common in current day cultural heritage research and communication. Nearly always, they represent large amounts of financial and intellectual investments, and rely on a wide range of data, people and interpretation processes. The documentation of the creation of such complex cultural heritage objects has been outlined in the widely accepted London Charter, but preservation of these objects in practice still has to overcome a range of issues. This short paper tries to provide an overview of the six most important issues and proposes some solutions and methodology to overcome them.

Six Issues to Deal with

The London Charter provides a very complete and well-structured framework to carry out documented 3D visualisation of complex cultural heritage (CH) objects, such as objects of art, man-made structures and historical landscapes. When focusing however on the practical implementation of the preservation of such digital visualisations, we need to deal with six major issues, which are:

- Lack of methodology to document and exchange 3D CH objects
- Lack of communication methodology
- Lack of stimuli to document and preserve
- Lack of long term storage and digital preservation strategies
- Lack of business models for reuse and exchange
- Lack of updating methodology

The **first issue** is still very basic: *we still do not have any tradition or adopted methodology or standard for how we document the creation of a 3D visualisation*. Although the London Charter[1] outlines very well the principles, we need a more practical methodology that can be adopted by the majority of people involved in 3D visualisation. There are already some initial guidelines, for example on implementing heritage visualisation in Second Life[2] or on general documentation of interpretation processes (paradata) in 3D visualisation[3] developed within the EPOCH Network of Excellence. But we need more good examples and best practices on how to take on such documentation activity, on what tools to use, on the workflow to follow. We need major involvement by the community to reach consensus that 3D visualisation and its documentation is a normal part of cultural heritage practice.

[1] London Charter (http://www.londoncharter.org/)

[2] The London Charter in Second Life (http://iu.di.unipi.it/sl/london/)

[3] EPOCH Knowhow book on Interpretation Management (http://media.digitalheritage.se/2010/07/Interpretation_Managment_TII.pdf)

But these guidelines need to deal not only with the lonely researcher that creates such 3D visualisations of complex cultural heritage, but also with teams, multidisciplinary and geographically distributed. In other words, *exchange* of such documentation and methodology to *collaborate* are essential elements in the practical implementation of a documentation and preservation strategy. This is clearly linked to the capabilities of the tools used. The InMan methodology[1] as developed within EPOCH used a wiki as medium for the documentation process, because of its discussion and versioning capabilities. Recent experiments for the 3D visualisation of the Abbey Theatre in Dublin[2] and the Etruscan Regolini-Galassi tomb[3] used a blog to record the interpretation process and related visualisation issues and to stimulate discussion and consensus creation amongst the group of 3D experts and a wide variety of cultural heritage experts. Although the idea of a peer review process of 3D models was coined already several years ago by Bernard Frischer in the SAVE concept[4], very little experience is already available on how exactly to implement such a review process, including source assessment, evaluation of 3D models, discussion amongst peers and improvement of the 3D model. Moreover, a peer review process should be based upon the London Charter and use metrics that reflect the London Charter principles. This still needs to be defined and a sufficient degree of consensus needs to be built on the methodology and implementation.

This brings us to the **second issue**, which is the *communication of 3D models of cultural heritage objects for collaboration, scientific publication and public use*. We need to make a clear distinction between these three goals, as they have different dynamics and requirements.

Within *collaborative research*, 3D models and their linked paradata need to be passed on from one expert to the other, for study, review and adaption. In practice, this turns out to be quite difficult, not only for technical reasons (ownership and knowledge of tools and 3D software) but also for organisational and psychological reasons. Our experience is that it works much better if one central person or team deals with the 3D models and paradata, while experts are consulted to contribute in their domain of expertise. This is also how the multidisciplinary team of Robert Vergnieux at the Ausonius institute[5] deals with 3D visualisation projects with great success. Using a blog is a useful instrument in this process, (see above), but our preliminary observations are that the blog needs to be private (limited to the research team), as experts are reluctant to contribute on a public blog as they see this as a kind of publication with final conclusions, while the contributions are ongoing research, of a volatile and progressive nature.

Scientific publication of 3D visualisation projects is quite common, but very few of these publications allow one to see the 3D results in 3D. As most publications result in a PDF file, we can use the 3D capabilities of PDF, which have matured significantly over the last 5 years. PDF is an open format, standardised as ISO32000-1 (an update of this ISO standard, i.e. ISO32000-2 aka PDF2.0 is in preparation). When authoring PDF documents, one can easily add 3D models into a publication[6] from a wide range of 3D formats. Other technologies, such as WebGL, start to be available to publish 3D online. The uptake of this simple approach however is hampered on one hand by the lack of 3D

[1] Ibid., see note 3 on p. 103.

[2] Abbey Theatre blog (http://blog.oldabbeytheatre.net/)

[3] 3D Visualisation of the Etruscan Regolini-Galassi tomb (http://regolinigalassi.wordpress.com/)

[4] Serving and Archiving Virtual Environments (SAVE) (http://vwhl.clas.virginia.edu/save.html)

[5] Ausonius Institute (http://www2.cnrs.fr/en/442.htm)

[6] Ni (http://www.nino-leiden.nl/doc/Annual Report NINO-NIT 2010 3D.pdf)

models in archaeological and historical research, and by the lack of education in the cultural heritage domain on how to use 3D in research and documentation.[1] Once there is a sufficient amount of 3D models created and used within the cultural heritage domain, there will be much more pressure to deal with proper 3D publication and digital preservation processes.

The most common purpose however for 3D cultural heritage data is *public use*. This means that a certain body of scientific results is used to show cultural heritage objects to the public in exhibitions, online or on TV. In this *public use* phase, the focus needs to be on the translation of those scientific results into a 3D visualisation that uses a certain medium (website, serious game, video, TV programme, etc.). Each medium has a specific language and the creation of results for public use from 3D cultural heritage objects needs specialists who master that language. For example, a director of a video can switch from 3D rendered images to animated drawings when the reliability of the visualisation of a certain object is too low. If we have for example insufficient data for making a reliable 3D model of a Phoenician ship, we still can show it as an animated line drawing (which also clearly conveys the message that we do not know these boats so well).

The CARARE project[2] is delivering cultural heritage objects (archaeology and monuments) to Europeana, partially in 3D. All of this 3D content does exist already but needs to go through a publishing cycle in which PDF is used for most content. Although part of the content will use 3DPDF[3] simply as a file format that can be displayed on every computer and OS, another part will need curated objects into which 3D is integrated in a document, with links from the text or the photographs to the 3D.

Many cultural heritage objects need to be reduced in resolution or complexity to be viewable online. Practice shows that too little effort is done to preserve the original high resolution data, while the low resolution public version is preserved, probably because it has much more visibility.

Efforts for establishing an *implementation framework* for 3D cultural heritage objects such as the Seville Charter[4], need to make a much clearer distinction between these three uses, and identify the different processes that are involved. In the research phase, the focus needs to be on collaboration tools to support annotation, discussion and consensus building. In the publication phase, the focus needs to be on optimal communication, linking the argumentation to the 3D models and passing on the 3D models for further research within the cultural heritage community. In the public use phase, the focus needs to be on transferring the 3D models and their relevant paradata to communication specialists, and using the right visual language of a certain delivery medium to convey the story that is told by these cultural heritage objects.

The **third issue** is how to *ensure that documentation and preservation of complex cultural heritage is made*. Practice shows that in most projects, over 90 % of the work goes into the analysis and interpretation of the data, while less than 10% goes into 3D modelling and texturing. Hence, failing to document and preserve the visualisation process results in the loss of at least 90 % of the invested money. For research projects that receive funding, documenting the visualisation process should be compulsory. Other

[1] Presentation on 3D technologies for Cultural Heritage in 3D PDF (http://www.faronet.be/files/bijlagen/blog/visual_dimension_v3.pdf)

[2] CARARE project (http://www.carare.eu/)

[3] 3D PDF – CARARE (http://carare.eu/eng/Resources/3D-Virtual-Reality)

[4] Seville Charter draft (http://www.arqueologiavirtual.com/carta/wp-content/uploads/2011/03/Sevilla-Charter.pdf)

projects within a more commercial context (for example commissioned by a museum to a company) can do the same as most or all of the budget for 3D visualisations is public money. In other words, we need to focus today on creating regulations that make documentation and preservation of digitally born cultural heritage objects a condition for funding or commissioning.

The **fourth issue** is how to *ensure that all these 3D visualisations, 3D models and their related paradata are stored for long term use*. This issue of course deals directly with several technical preservation issues, such as the file format of the 3D models and the documentation of all related files (textures, bump maps, etc.), for which strategies are in hand. But technical preservation issues are only a part of the problem. Although universities and companies can exist for a long time, research teams and company teams are normally quite transient. This means in practice that universities or companies are not the right place to store complex cultural heritage objects. In our opinion, storage in a repository at the national level should be compulsory (see for example the ArcheoGrid repository[1] at the national level in France). Storage at such a repository should be subject to a selection and prioritisation procedure on what to preserve and what to let go, to limit the cost of registration and storage. Ownership and IPR should be clear at least at the moment of storage.

The **fifth issue** is the *creation of a model for possible reuse*. It is conceivable that 3D cultural heritage objects should be available for free in low resolution to the public (for example in Europeana[2]) but can have a paid use for high-resolution versions. Museums can use and exhibit digital museum objects from other museums, digital publications can incorporate high-resolution 3D digital objects (for which royalties will be paid, just like for professional photographs today), even film companies and game developers could pay significant fees to use scientifically correct 3D models of historical buildings and objects. The V-MusT.net project[3] is developing a business model and practical implementation for exchange and reuse of digital museum objects and virtual environments.

Finally, the **sixth issue** is *how to keep 3D visualisations of cultural heritage objects and their related paradata up to date*. Nearly all digitised cultural heritage objects are static as they represent the physical object as truthfully as possible. Practice however shows that 3D visualisations are not static at all, they are based upon sparse data, hence they change because of the availability of new research, better excavations or new insights in the use and meaning of objects. The ideal situation would be that any 3D visualisation research by a certain team could be taken up by any other team that improves and complements the results of the first team. If that ideal situation were present, updating 3D visualisations would be quite a natural thing to do. However, this ideal situation is still a distant dream. We can bring that ideal situation a bit closer by dealing with all the issues described above. But still, there will be an important issue of cost, as we need to balance the available resources. What is the use of documenting the finest detail of a 3D visualisation project if that documentation is never used again? In other words, we still need to find out what the optimal amount of documentation and preservation should be, so that long term overall costs are minimised. This is still uncharted territory and needs more research and practice.

[1] ArcheoGrid (http://archeogrid.in2p3.fr/)

[2] Europeana digital library on culture (http://europeana.eu/)

[3] V-MusT.net Network of Excellence (http://v-must.net/)

Conclusions

The conclusions of this short paper are quite simple: the London Charter provides an excellent framework for the digital preservation of complex cultural heritage objects, but we need to focus now on putting the principles of the London Charter into practice.

This means that we need to collect *best practices* by analysing existing projects to find out which approaches do work and why.

This also means we need – based upon the conclusions of these best practices – to define *optimal workflows and specific requirements*, so that documentation and preservation of complex cultural heritage objects becomes an integrated part of cultural heritage practice. Crucial in this process is the definition of *quality* for the documentation, communication and preservation steps, as described above, and taking care of an appropriate balance between resources, results and the impact of those results.

Finally, we need to realise the *uptake* of such workflows and requirements, first of all by integrating them as soon as possible into the *curriculum* of students in the cultural heritage domain such as archaeology, history, anthropology, monument care and museology. Another major step into the uptake of such documentation, communication and preservation strategies can be the *Competence Centres* that involved European projects such as V-MusT.net[1] and 3D-COFORM[2] are setting up to support cultural heritage institutions and their partners.

And why not look into *expanding the London Charter* with clear guidance on these processes, based upon a wide consensus in the cultural heritage domain, so that it can acquire the status of a real Charter that governs documentation, communication and preservation of digital cultural heritage objects?

[1] Ibid., see note 3 on p. 106.

[2] 3D-COFORM project (http://www.3d-coform.eu/)

Digitising Mask Miniatures: A Life beyond a Museum Display

Martin Blazeby

King's Visualisation Lab, Department of Digital Humanities, King's College London, 26-29 Drury Lane, London WC2B 5RL, UK

Abstract. The Body and Mask in Ancient Theatre project, led by King's Visualisation Lab, King's College London, funded by the Arts and Humanities Research Council, concerns the body and mask in ancient drama – specifically in terms of intercultural performance and perceptual experience. As well as creating an online archive of digitised representations of ancient mask artefacts, the project has recreated a selection of iconic examples in 'full size' as wearable masks and recorded outcomes of staged performances which can be played back and interrogated for scholarly theatre research. Using various computer technologies it addressed fundamental questions relating to the conditions and actualities of ancient theatre. What can be inferred of the actor's technique and use of mask and body? How were these phenomena experienced in the differing theatre spaces of Greece and Rome?

Introduction

It is acknowledged that masks were once a fundamental feature of ancient Greek and Roman theatre; however, it is unfortunate that these masks do not survive in physical form to this day. Masks used in theatrical performances were in most cases constructed using perishable organic materials such as wood, natural fibres, shell, animal fur, and hair etc. All of our current knowledge can be derived from ancient literary descriptions including play texts, and our visual understanding of ancient masks can be learned from surviving iconography such as sculptures, mosaics, wall paintings and depictions on vases.

Mask miniature sculptures that were made from metals, terracotta and marble have survived and these miniature artefacts, although not masks themselves, provide the best examples in three dimensional physical forms as to what 'full size' wearable masks may have once resembled. These mask miniatures are now located in abundance in Museums across Europe and despite varying in sizes, materials and site provenance, they often share commonalities of features. Theatre scholars refer to Pollux and his *Onomasticon* (IV, 143-154), as a way to identify these mask miniatures as Pollux provides a fairly comprehensive list relating to comedy and tragedy. This classification also includes various sub-divisions but mainly focuses on the main character types of Old man, Young Man, Female and Slave.

As a way to further understand and expand our knowledge of ancient masked performance practices the project needed to capture the physical three dimensional attributes of mask miniature artefacts using available digital technologies, and re-create them in such a way that they could be worn at 'full size' by acting professionals already familiar with masked traditions. It would be impossible to digitise all mask miniatures, therefore a selection of masks were sought by the project team in order to represent a range of character types, including age and gender. The actors then performed scenes from ancient dramas enacting these various character types whilst wearing the 'full sized' masks and appropriate costumes. Each scene was then recorded with multiple view points using various technologies as a way of documentation and also to situate the results within previously modelled 3D theatre environments.

The resultant performance footage when combined with a virtual theatre context will hope to raise issues relating to performance spaces and their history of general interest within theatre studies, including the difficulties of working with masks in large semicircular auditoria, and the differing experiences of mask and body for spectators at different angles and distances to the stage. The output footage will also aim to form an objective basis to investigate perception of masked performance, and to isolate how specifics of shapes, tensions and body configuration are read both by those acquainted with different performance traditions and those unfamiliar with them.

Technology Overview

Various technologies were used to digitise mask miniatures including 2D photogrammetry, portable 3D laser scanners and fixed based 3D laser scanners. The use of these technologies was essentially dictated by individual museum policies and protocols relating to accessibility, curatorial supervision, including artefact handling and an understanding of the technologies involved. In some cases verbal explanations and signed agreements were necessary to persuade museums that laser scanning was a non-intrusive technology that did not damage artefacts particularly in regards to surface colouration. The most difficult aspect to scanning on site museum collections was organising visitations corresponding to public closure days and curator availabilities. The majority of the mask miniatures sort by the project was on public display and locked behind glass cases, therefore the minimum of disruption was paramount to curators, security and the project team. The Musée du Louvre, Paris was an example where project members had to pre-select in advance a handful of mask artifacts from a wish list off site, the actual digitisation on site using a portable Minolta 3D scanner was then governed by curatorial supervision that involved extreme care of artifact handling which dramatically slowed down the process. The Ny Carlsberg Glyptotek museum, Copenhagen allowed the team to scan using the same portable Minolta 3D scanner from a pre-selected list with unrestricted supervision and also a freedom to touch and manipulate artifacts during public accessible days. This in turn led to an unrushed but quicker turnaround of total data capture than that of The Musée du Louvre, and enabled additional artifacts to be processed during the available timeframe. In the case of the British Museum, each mask miniature chosen by the team was individually transported one at a time by means of a taxi along with curator just a short distance to the department of Geomatic Engineering at UCL. Here the artifact was digitised using a fixed bed Arius 3D scanner one of the highest quality 3D scanners in the world and operated by experienced in-house staff.

The post production of scanned data was done off-site using a myriad of 3D software packages, the majority of which are exclusive to manufacturer and data capture method implemented. Autodesk 3D Studio Max was the core software then used to correct digitised data and convert files into STL (Stereolithography Interface Format). These files were then used to print from in the way of rapid prototyping, in order to create a positive mould so that a cast can be made to form a 'full size' wearable mask. 3D Studio Max was also used construct the accurate architectural representations of ancient Greek and Roman theatres, these virtual environments had previously been created by the KVL team for various projects and a selection would form the contextual settings for masked performance analysis.

Adobe Photoshop was an integral tool in creating hypothetical 2D representations of the 'complete' mask in terms of painted or decorated features. Various mask types were created using Adobe Photoshop and these were then used to directly overlay onto the relevant 3D mask artefact in way of 2D texture maps. The resultant decorated mask was then exported from 3D Studio Max as a real-time 3D web viewable model using the TurnTool freely available plug-in.

Two main technologies were used to record the staged performances of actors wearing the reconstructed masks, chromakey video recording using Reflecmedia Chromatte fabric backdrops and full body motion capture systems. Adobe Premiere enabled edited sequences of video footage to be combined with composited 2D renderings of 3D reconstructed Theatre structures as backdrops. Motion Builder software allowed raw motion capture data to be applied to a bipedal animation rig (the digital equivalent of bones for a CGI character), the results of which were directly imported into the 3D Studio Max files of selected reconstructed theatres.

The project utilises the following pipeline processes and methods:

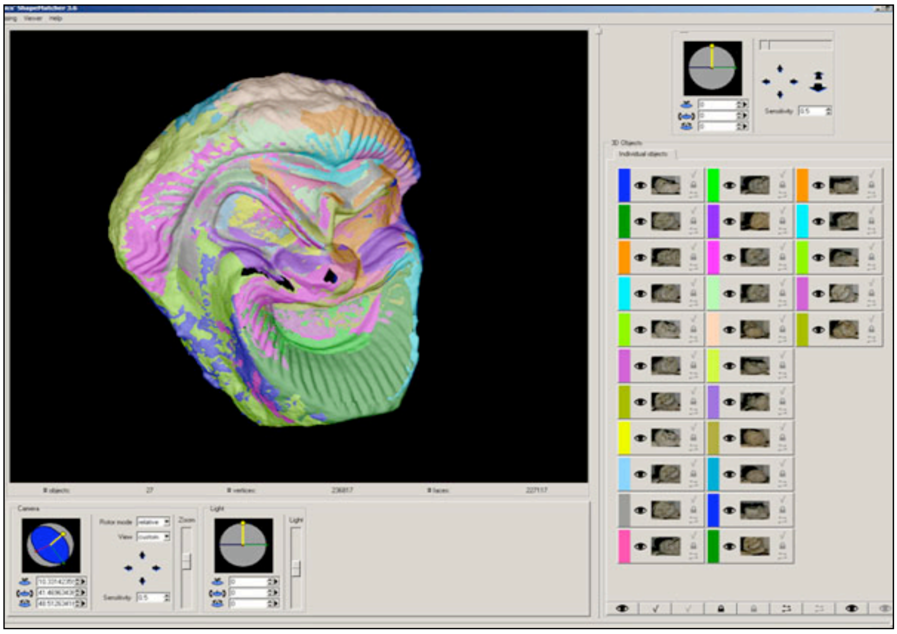
- 2D photogrammetric scanning and 3D (portable and fixed bed) laser scanning of mask artefacts.
- Post scanned editing to correct, modify and reconstruct digitised mask artefacts.
- Hypothetical 2D decorated mask reconstructions in Photoshop.
- Creation of replica artefacts by 3D rapid prototyping.
- Hand constructed ' Full size' masks by a professional mask maker.
- 3D visualisations of Greek and Roman theatre environments.
- Chromakey video recording & full body motion capture of masked actors during staged performances.
- Post production editing of video footage & animations of motion capture sequences.
- Web based catalogue of digitised masks and dissemination results.

Photogrammetry

Photogrammetry is a low cost, portable and non-intrusive technique of recording 3D objects by means of photographs or video. The main components consist of a slide projector, a slide containing a fine grid, a calibration box and a camera or video recording system with tripod. The process of photogrammetry is as follows: the slide is projected onto the area where the artefact will be positioned, the calibration box is placed in situ and a photograph is taken from a fixed position. This calibration photograph is crucial, as it is required to calculate the distances and angles between camera, projector and object. When this is complete the artefact can be photographed with the grid projected onto the surface, it will be necessary to rotate the object in order to record all angles covered by the projected grid. The amount of photographs required will vary depending on how fine the grid system is and how complex the artefact is.

Using photogrammetry software (Shape Snatcher) the calibration photograph is loaded and the relevant photographs are imported. The software interprets the two dimensional photographs containing the grid projected on to the surface of the artefact into three-dimensions. Each segment of the artefact as photographed from different angles is then saved separately and stitched together as one 3D mass. This process can be time consuming as manual input is often necessary as automatic alignment can often lead to miss matched segments. When successful the produced results are of sufficient quality for use within the project, however, this approach to data capture may not produce data of required granularity suitable for other projects that demand a higher resolution of detail. Shape Snatcher software can then export the completed artefact as a polygonal mesh object such as obj format which can then be imported into other 3D software packages without any difficulties.

The significant drawback of using photogrammetric data capture is that it relies heavily on the stability of both projector and camera, if either of these has accidentally been moved after the initial calibration image is taken then the subsequent data is at best questionable. This factor is based on the software interpreting the precise positioning of the camera and projected grid as recorded by the calibration image, every photograph taken thereafter is matched in accordance and calculated as a three dimensional segment, any movement of the projected grid mid session will result in misalignment of these segments during the stitching process. This can be problematic especially when carrying out post production stitching off site and far away from a museum collection, in certain cases a partial capture can be recovered and sometimes an entire data capture session will need to be abandoned and will have to be reacquired if allowed.



objects 27 # vertices 236817 # faces 227117

Rotor mode: relative

View: custom

Sensitivity: 0.5

Figure 1. Shape Snatcher screen shot showing stitched photogrammetric segments of a slave mask artefact recorded at the Fitzwilliam Museum, Cambridge.

Laser Scanning

For the purpose of the project we have used two types of laser scanners, a fixed based Arius 3D laser scanner located at the department of Geomatic Engineering, UCL and a portable Minolta laser scanner loaned by the Archaeological Computing Research Group, University of Southampton. Both of these laser scanners are non-intrusive and measure highly accurate levels of detail, characterising each measurement point according to its colour and location. Each measurement point is recorded by three geometric values as XYZ and three reflectance values as RGB collected simultaneously from the target surface.

The Arius 3D scanner is a very large non-portable machine allowing for data capture to only take place at UCL. This requires a high level of forward planning and management as the artefact required for scanning will inevitably have to be removed from its museum location and transported off site to the scanner. Security and safety is paramount in an operation like this, usually involving the time of both the scanning staff and of museum curators who will accompany and protect the artefact at all times. Scanning is recorded as point cloud data and the resolution captured is at sub ***millimetric*** level which is way more than sufficient for use within the project. This level of capture, although not necessary for our requirements, provides an archive of significant resolution perfectly suitable for conservation and archiving purposes, therefore potentially enabling further research and interrogation for use beyond the scope of the Masks project.

The Minolta portable laser scanner, unlike the Arius 3D scanner allows for far greater portability, allowing scanning onsite at various museums. This however, requires permissions and unrestricted access to specific museum collections and archives. A great deal of planning is essential when dealing with European collaborators and an explanation of the scanning processes to officials is a prerequisite to those who may otherwise not be technologically understanding of the process involved. The Minolta scanner is robust for transportation and once setup and operated by its handler, the scanning process is quick, although not as detailed as the Arius 3D scanner, it is still accurate and precise.

The scanned data created by both technologies is then processed by proprietary software to stitch component segments together, similar to that used in the photogrammetry. However, these processes were more automated with self-alignment considerably faster and more accurate. It is worth noting that despite the point cloud data recorded by the Arius 3D as being extremely high in resolution and precise in colour information, it needed to be converted using Rhino CAD software into polygonal mesh data and reduced in granularity.

3D studio Max

Each resultant digitised artefact required varying degrees of post-production modification in order to correct and repair areas of 'bad data'. This may seem astonishing considering the sophisticated standard of technology utilised. However, factors contributing to 'bad data' were not necessarily the result of the technology itself, but more due to elements caused by artefact surface properties. For example, shiny regions where the laser beam could not detect the correct surface depth; also, areas containing deep recesses know as occlusions, such as eye sockets and nostril holes where the laser beam or photogrammetry projected grid could not penetrate fully. Other factors relating to the various stages of import, export, file conversion and mesh reduction also add to underlying levels of data corruption.

Using 3D Studio Max as the core software package, each completed mask scan was subject to close inspection to determine the level of correction needed. In some cases no changes were necessary, in others, erroneous mesh data was removed and missing or inverted polygons re-created or adjusted. Several masks would have required substantial working and hours of labour in order to fill in large holes and missing features, it was however, deemed inappropriate as vast amounts of correction bordered on degradation and would corrupt the integrity of the original artefact.

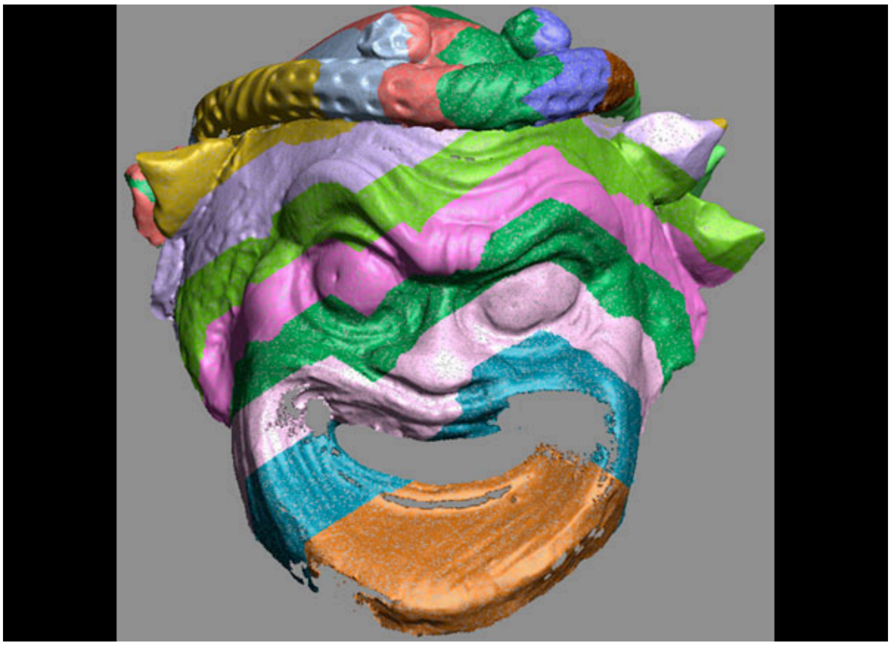


Figure 2. High resolution laser scan passes of slave mask artefact from the British Museum. Sections missing from around the mouth were too significant to re-create during post production.

After each digitised mask had been edited and finalised in 3D Studio Max, a selection relating to different genres were chosen to be used as the 'full sized' versions. These masks were subjected to a process of virtual scaling in order to correspond with 'real life' physical proportions of a human head. Guidance was sort regarding measurements of an 'average' head size and spacing between facial features etc. and a 3D head template created. The digitised masks were then scaled approximately to fit the overall dimensions of the 3D head template, but the head was used as guidance principally to locate and position the eyes and mouth openings of the masks.

It is important to remark that the masks artefacts are miniatures by their very nature and by scaling them up in size (in some cases up to 500 %) would not always correspond successfully with 'real life' proportions. This can be attributed to the fact that the facial proportions of the miniatures relate to exaggerated mask characteristics and probably not intended to adhere to any human proportions in any case, also due to their size, any flaws relating to craftsmanship or archaeological condition will be amplified when blown up to 'full size'. One such example was demonstrated when an excellent correlation was achieved between eye, mouth position of what seemed to be a good mask candidate for 'full size' use, but the physical space within the inner mask area around the jaw would not be sufficient enough for the movement of an actor's mouth and chin during a performance.

The virtual scaling of the digitised masks was a hit and miss process, however, it soon became apparent that the masks that achieved good positioning of eyes and the mouth opening, but were slightly larger in terms of head height and width would be best suited. This is because padding could always be applied to the mask during the actual construction phase of the project.

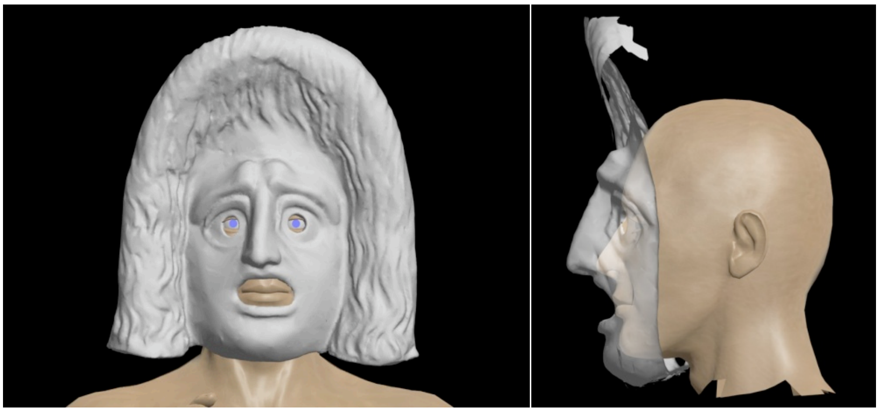


Figure 3. Images showing digitised mask and 3D head template positioning with mouth and eye alignment.

Rapid Prototyping

With the masks now scaled and approved by project members, they were exported from 3D Studio Max into STL format and sent to the Hothouse Centre for Design, a company specialising in fast turnaround services for rapid prototyping printing. It was agreed that one mask was to be printed as a test case as a way to gauge that the 3D head template was adequate enough in size and that the data was of sufficient quality. Rapid prototyping is very expensive for one off print jobs, therefore vital that each mask was checked several times throughout the scaling and exporting process, any miss calculations would be extremely costly for the project.

Having received the first mask printed with a robust lightweight material formed with a plaster and gypsum compound, it was worn and tested by various people within the project and the go-ahead was given for the rest of the masks to be printed. Due to rapid prototype constraints at the Hothouse Centre for Design (the build chamber is 25.4 x 35.5 x 20cm cubed); masks that were larger than the build chamber would have to be printed out using multiple segments. One mask in particular was slightly problematic, mask MYR349, scanned at the The Musée du Louvre, Paris, contained a large hair like structure known as an *onkos*. In order to print this mask in its entirety it would have to be split into three sections then, after printing, glued together to form a complete structure. This method was considered too costly and it was thought unnecessary to have the whole hair piece reproduced in the print process. However, this came after some deliberation relating to the hypothetical material make-up of the hair: Was the *onkos* supposed to represent a 'solid' structure or 'real' hair, or a combination of both?

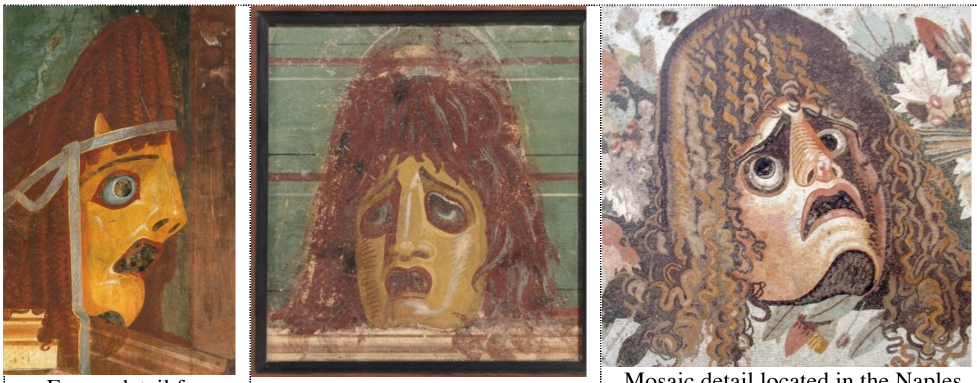
It was agreed by project members that the project was more concerned with creating 'wearable' masks for performance based on ancient miniatures which themselves were based on masks, and not necessary about creating exact 'full size' replicas of such artefacts. It would however, still focus as best as possible on an exact impression of the facial characteristics as recorded by each original scan, but any features that would fit outside of the print area would be created as representational by hand afterwards by the professional mask maker.

Virtual Decoration

Before the rapid prototyped masks could be sent to the mask maker and turned into wearable versions, fully worked up colour designs needed to be created for each individual mask type. These designs, although hypothetical, were based on the written descriptions by Pollux in his *Onomasticon, and* comparanda of masks that are depicted in frescoes and mosaics. There seemed to be some correlation between the surviving colour depictions that were thought representational of the masks used in the project and related descriptions by Pollux in terms of character types, facial colouring and expressions etc.

One good example used within the project would be that of mask MYR349, the young man with *onkos*. Three particularly good images portraying this type of mask were used as comparanda; one of which depicted in a large fresco from the Villa of Oplontis, Torre Annunziata, Italy; is excellently persevered on the east wall of room 15. The other two images located in the Museo Archeologico Nazionale, Naples, one of which is a fresco segment originally from Region IV Pompeii and the other a mosaic originally from the House of the Faun, Pompeii.

What can be established form these graphic depictions is a common pattern of hair style, colouring and facial shape all of which are also described in detail by Pollux. The hair is formed out of many curls or ringlets, the face colouring has a yellow hue and the facial features formed using similar stylised markings. The area around the eyes has a prominent red ring carefully curving around the shape of the eye, the eyebrow appears to point upwards towards the forehead and the lips pronounced in red contouring the downwards shape of the mouth. It is also worth noting that these depictions show that the underside of the chin is cutaway presumably to enable ease of fit to the actors' face, all of which show clear and similar characteristics of a tragic *onkos* type mask.



Fresco detail from Room 15, The Villa of Oplontis

Fresco located in the Naples Museum. Originally from Region IV, Pompeii.

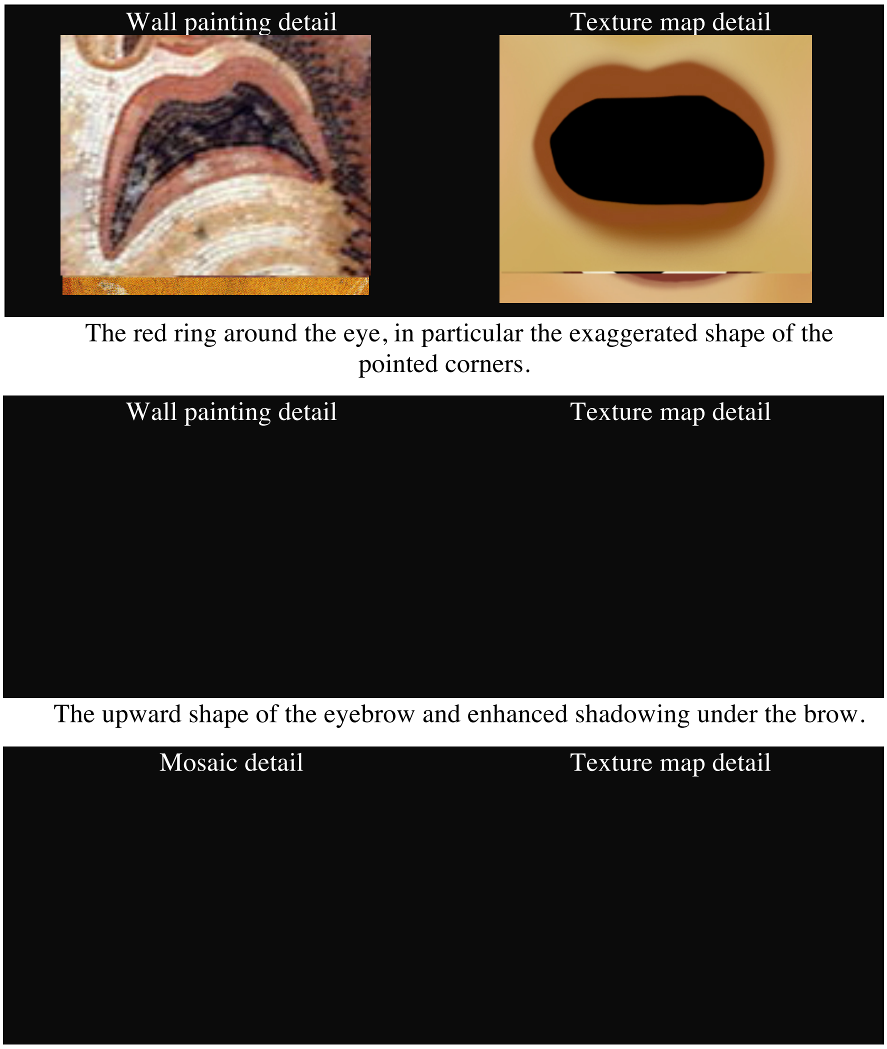
Mosaic detail located in the Naples Museum.

Figure 4. Examples of *onkos* type mask.

The first stage of the virtual decoration process was to take an orthographic rendered image of the 3D mask in 3D Studio Max; this was taken from the front, facing the viewing plane in order to omit any distortions in perspective and angle. The resultant rendering was used to form a base layer in Adobe Photoshop providing a template on which to build up precise layering of painted textural information. This methodology is such that it allows for an exact mapping to take place between the virtual decorated images created in Adobe Photoshop and their alignment back onto the 3D masks in the way of planner UV texture maps.

The decoration process is subjective and somewhat experimental, relying on various painted stokes and colour variations to be continually made in Adobe Photoshop and then reloaded into 3D studio Max in order to build up the required effect as viewed in three dimensions. This process is necessary to enable the precise matching of paint markings around intricate areas of the face to the corresponding areas of the mesh object. An imprecise paint stoke could sometimes lead to serious distortion as it is mapped on to the contours of the object.

Significant key arrears of facial characteristics for replication include:



The red ring around the eye, in particular the exaggerated shape of the pointed corners.

The upward shape of the eyebrow and enhanced shadowing under the brow.

The fullness and downward shape of the lips.

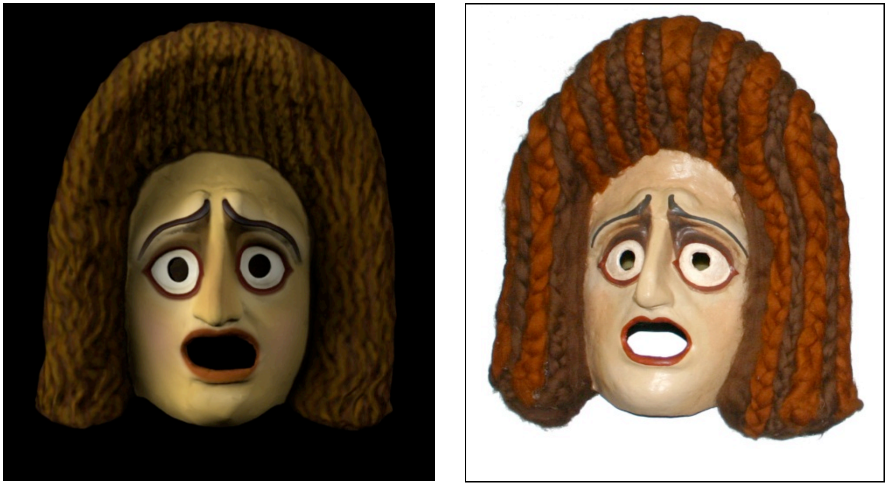
Figure 5. Key arrears of facial characteristics

‘Full Sized’ Performance Mask Making Process

On receipt of the rapid prototyped mask made from plaster-resin powder, Dr. Malcolm Knight an experienced mask maker at the Scottish Mask and Puppet Centre, embarked on a complicated and skilled process to construct a version that would be capable of being worn during a performance. Firstly, a surrounding outer shell was formed around the mask using clay, filling in any holes to create a fully sealed unit. The clay shell was then removed from the rapid prototype mask and covered with a silicone infill forming a negative mould from the positive shell. A fibre-glass coating was then applied and inlaid with layers of torn wool paper ('carta lana') soaked in rabbit size hot glue until a uniform thickness was achieved of about one-sixteenth of an inch. The wool-paper mask is left to dry for a period of a couple of days and then is released from the silicon mould and checked for any irregularities or distortion.

The newly formed positive mask is then mounted on a plaster head bust so that a rear section (helmet) can be constructed in clay. Layers of resin cloth (celastic) were then applied to the clay helmet allowing for seamless bonding between the wool-paper mask to take place forming a unified structure. It is worth noting at this point that some of the mask miniature artefacts contained a rear helmet section, and these were digitised during the scanning process and printed out during the rapid prototyping process. However, more often than not, these rear sections when tested would not fit the physical attributes of a human head. Therefore, it was acknowledged that all helmet sections would be custom made regardless of their original existence. The resultant mask with combined helmet is carefully detached from the plaster bust and the clay, at this stage the openings in the mask for eyes, nostrils and mouth are then cut out using a selection of cutting tools and the surface of the mask and helmet is sanded smooth.

The 'full size' mask is then decorated in accordance with the 2D rendering of the virtual mask. The face is painted using a white primer base coat and then slowly built up with acrylic colours to achieve and enhance the necessary character features. Materials such as sheep's wool and rope are used to create hair and are carefully applied in such a way as to mimic to the original artefact and digital version of the mask. The finalised mask is then protected using varnish and padding applied to the interior so that areas around the top of the head, cheeks and chin can provide cushioning and comfort for the actor. To Quote Knight "This process of mask-making is relatively low-tech and the essential stages of casting, building, cutting, sanding, painting, finishing and padding have been practised for centuries by all *skeuopoios* ('makers of kit' as the ancient Greeks called them)."



Finalised 3D rendering of mask MYR349 with applied texture map.

Completed 'full size' performance mask MYR349.

Figure 6. Example of a completed performance mask and its finalised 3D rendering

Performance Workshops

A major research element and central to the project concerns the recording of actor(s) wearing the reconstructed masks during a pre-set series of workshops. It is understood from masked performance as it is practised today, that the wearing of a mask during a staged performance influences the body in terms of movement in a given space. Therefore, it was essential to learn how the movement of the body is expressed and informed by these Greco-Roman replicas as would have been experienced by spectators within the context of ancient theatres of varying sizes and structure. The workshops featured the performer(s) and styles of Akira Matzui (Kita School of Classical Noh Theatre), Angelo Crotti and Romans Suarez-Pazos (Italian Commedia dell'arte) and Wayan Dibia a professor of dance at the Indonesian Art Institute, Bali. All of these practitioners carry a wide cross-cultural relevance and expertise in masked performance traditions.

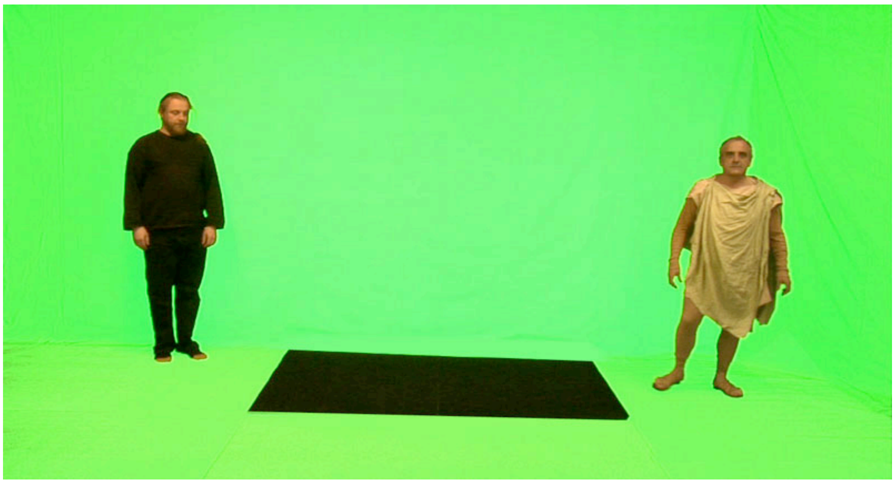
Instead of performing within real theatre spaces, the actor(s) had to perform within 'imagined' spaces as they were recorded using chromakey and motion capture technologies. Virtual ancient theatre settings were carefully selected and the relevant 3D models of these spaces were prepared, ranging from temporary stages, small performance spaces, medium sized roofed odea and large sized open theatres. The actor(s) analysed these theatre spaces in advance so that they could prepare specific tailored movements relating to the scale of projected auditorium.

Chromakey technology was chosen as a way to document the body's nuances through photo-realism, and would also record the movement of clothing and sound. However, the output can only be viewed from a fixed perspective as recorded by the camera, whereas, motion capture would help to investigate less subtle characteristics focusing on movements generated through larger spaces and played back as real-time motion from any angle. It was felt necessary for the project to use both technologies and therefore the actor(s) performed two versions of the same sequence.

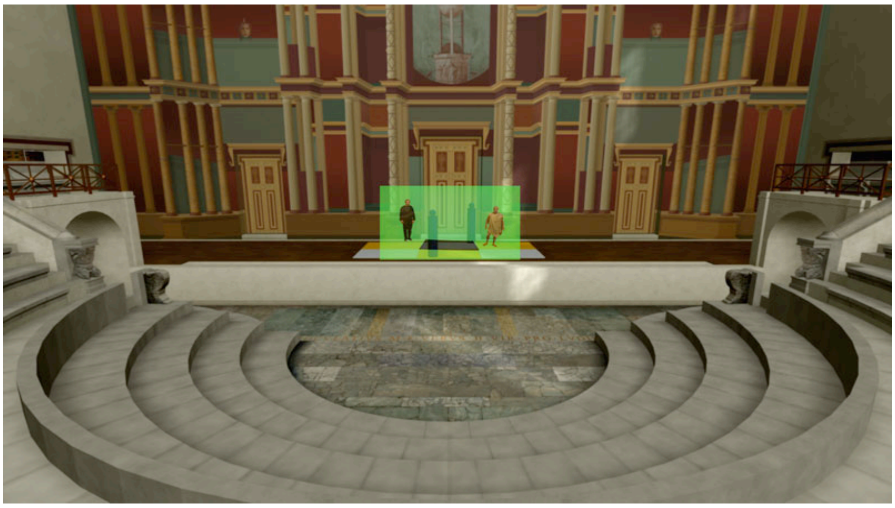
Chromakey Video Recording

The location of the chormakey video camera within the designated studio was important; the intention was to create viewing angles that would correspond with sightline specific viewpoints from selected theatre spaces. The actors would then perform in front of chromatte fabric backdrops (similar to green or blue backgrounds used by weather forecasters), which would later be replaced with virtual theatre backdrops.

Working on-site, the dimensions of the physical performance area of the studio along with the heights of each actor were plotted and referenced in each of the 3D theatres to be used as backdrops. A calibration still image was taken of the actor in the performance area using the studio video camera; this was then compared by overlaying a calibration still image taken from the virtual spectators' position within the 3D theatre model. The two calibration images needed to match as precisely as possible with each other, for this to succeed, the angle and projection of the studio camera had to coincide with the angle and projection of the virtual camera; this then would ensure an exact alignment between the two perspectives and exact proportions relating to the height of the actor(s). This calibration process sometimes took several attempts for a like for like comparison, resulting in fine positioning of the studio camera and virtual camera.



Chromakey calibration image taken eight metres from the performance area.

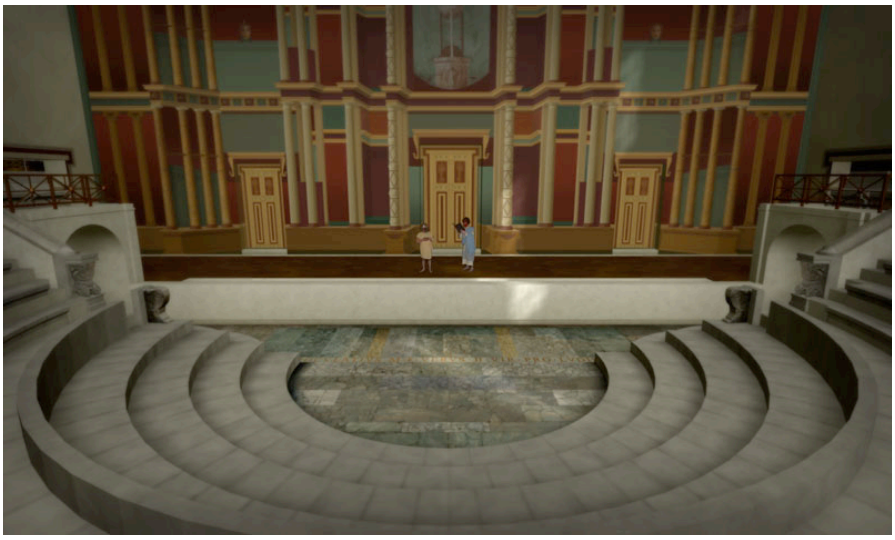


Calibration image scaled to match a virtual viewpoint as seen from 15 metres away.

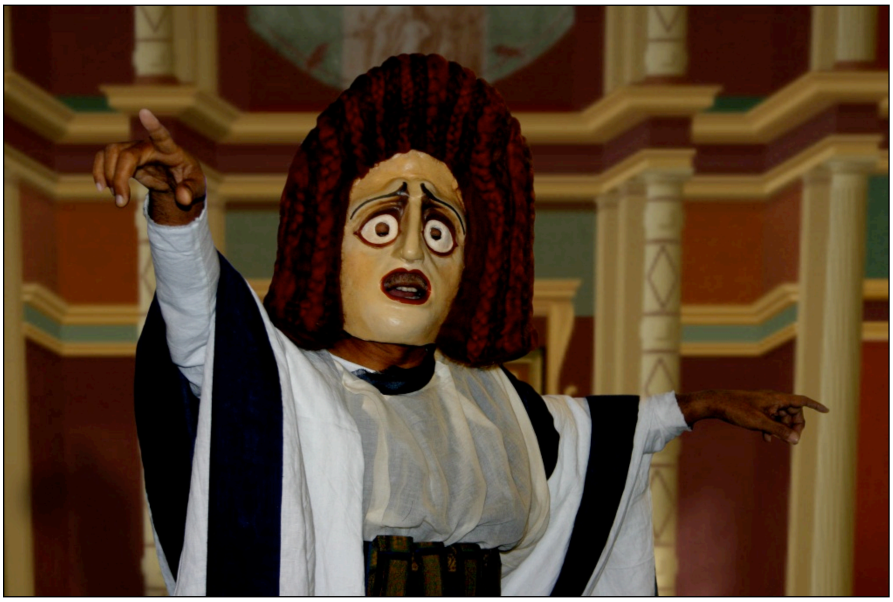
Figure 7. The calibration process

Due to limitations of the studio, the camera had a maximum recording distance of around 8 metres; therefore, it was impossible to match accurately a virtual camera positioned in a large theatre. For example, the maximum distance calculated between that of an actor and camera viewpoint in the Theatre of Pompey is in the region of 70 metres. It was formulated that the only viable solution was to disregard perspective for larger theatre contexts and to concentrate on angle and trajectory instead.

The resultant video footage was then subjected to post production editing using chromakey software. The chromatte background is removed (keyed-out) and replaced with the view point specific virtual theatre backdrops. The masked actor(s) then scaled according to the calibration template created for each theatre. In general, the finalised composited sequences were deemed successful; they enable pseudo realistic renditions of precise audience sightlines which otherwise would be difficult to visualise for scholarly analysis.



Composited actors prepared for a medium sized theatre setting.



Composited actor prepared for close range viewpoint.

Figure 8. Composited actors from different viewpoints

Motion Capture Recording

Motion capture supports the recording of movement in a given space similar to chromakey video technology; however, skeletal data is obtained tracking specific positions of the body. The motion data output, unlike 2D linear video, can be viewed from any position and angle in three dimensional space through user controlled play-back. This clearly offers distinct advantages of bodily analysis relating to velocity, mass and muscle forces but the results are less immediate requiring levels of post-production processing.

Two types of motion capture systems were employed in the project, Gypsy exoskeleton suits and Vicon optical tracking. It became immediately evident during the first workshop involving Akira Matzui that the exoskeleton suits inhibited movement due to their construction and would provide inadequate results. The Vicon system generously provided by Coventry University offered a better solution as it enabled far less restriction of movement due to external optical infrared recording. The actor wears a motion capture suit with tiny markers attached to key locations, cameras located around the laboratory are then used to triangulate the markers at a capture rate of 250 times per second with incredible accuracy.

The raw motion data can be analysed in simple skeletal form or can be processed into skinned characters using Motion Builder software with clothing and masks attached. This is a skilled process usually associated with the games and animation industry; fortunately Wong Choi an expert in mocap editing and kinetics at the School of Science and Engineering, Ritsumeikan University, Japan, assisted the project with the creation of virtual actors which were then situated within the reconstructed 3D theatres and reviewed from multiple sightlines.

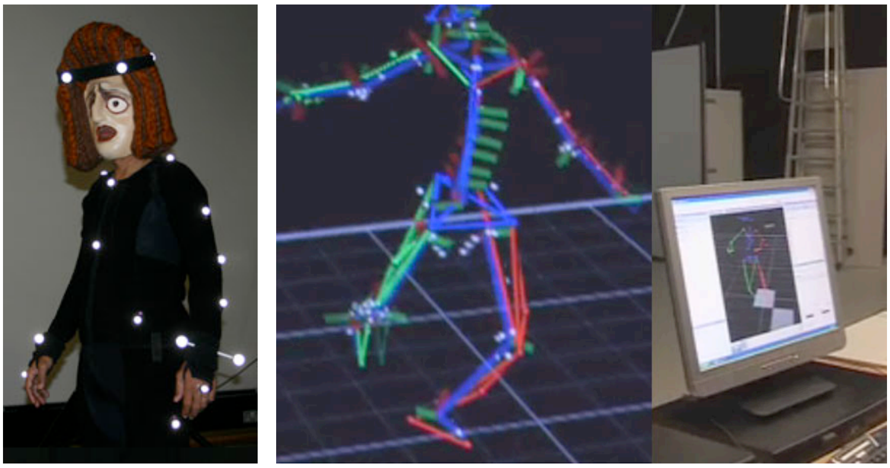


Figure 8. Vicon optical system with onsite skeletal analysis.

Summary

The project has undertaken research using state of the art technology to investigate whether masks miniatures contain sufficient enough information to be scaled up in size and worn by experienced actors during practice based workshops. Without these surviving artefacts it would be impossible to recreate ancient masks as faithful three dimensional forms for scholarly study and interrogation. Therefore, it was paramount to ensure that the digitisation process and methodology used to capture the masks was of great importance, enabling the acquisition of precise data to be recorded using the best technology and tools available. In most cases the resultant scaled masks proved to be successful which was surprising considering the original sizes of the miniatures and the intricate carving techniques used to construct them.

The scanned mask data, as well as providing the core content to facilitate the various stages of the project has been archived, documented and disseminated online. This unique database presented as a single resource and referenced by respective theatrical character types has brought each mask to life in a way that could never be experienced in a traditional museum context. Many of the original artefacts are distributed across European collections and have limited accessibility; some of these are located behind glass cases and can only be viewed from the front, some are too small to be appreciated and other locked away due to their fragile and precious nature. Therefore, the scanning of these masks offers advantages in providing research through remote access, digital preservation for future reference, and may also contribute knowledge to other research communities, specifically those interested in ancient history, theatre and sculpture.

The practical experiments using the 'full sized' masks have also given the project an insight into ancient performances, of significant interest was the discovery derived from the workshops relating to the effectiveness of the mask in audience perception as viewed over a range of distances. Although the masks created for the project were relatively detailed in physical form and made expressive through means of painted character features, they appeared to only communicate successfully at close range. This discovery prompted questions beyond the intended scope of the project primarily – *were masks generated for specific theatre venues, in terms of decoration?* If so, would an indoor theatre require brighter colours or conversely would large outdoor venues require oversized masks with cruder painted features*?*

As a way to evaluate this hypothesis, the project sought to transform a selection of masks originally decorated by Knight to enhance individual character traits by painting over existing features with contrasting colours and bolder markings. The masks were then worn by the actor(s) and recorded once again, coinciding with identical scenarios as the unaltered masks in order to match like for like comparisons. It was clearly evident that the re-painted masks had a much greater impact when viewed from the virtual setting of a medium sized theatre, but still had little effect when viewed from the rear seats of a large theatre such as the theatre of Pompey. When viewed at close range the perception of the re-painted masks had changed significantly, the exaggerated features had detracted from the original aesthetics thus creating masks that looked more sinister and grotesque in appearance.

No credible conclusion can be derived from these experiments; however, the results suggest that there is a distinct relationship between the body and the mask as experienced in different theatre spaces. It is plausible to suggest that some masks were only suited to particular venues and that other masks were custom made for individual performances; or perhaps ancient masks were constantly modified just as undertaken in the project so as to communicate effectively from one theatre space to another.

References

Museums and collections made available for scanning of mask miniatures and related artefacts:

- British Museum, London
- Museo Archeologico Nazionale di Napoli
- Ny Carlsberg Glyptotek museum, Copenhagen
- Musée du Louvre, Paris
- Thorvaldsens Museum, Copenhagen
- Soprintendenza archeologica di Pompei
- Deposito Antiquarium Ercolano

Project website: http://www.kvl.cch.kcl.ac.uk/masks/

Digital Preservation Strategies for Visualisations and Simulations

Janet Delve[1]**, Hugh Denard**[2]**, William Kilbride**[3]

[1] Future Proof Computing Group, School of Creative Technologies, Eldon Building, University of Portsmouth, PO1 2DJ, UK

[2] Digital Humanities, King's College London, 26-29 Drury Lane, London WC2B 5RL, UK

[3] Digital Preservation Coalition, Innovation Centre, York Science Park, Heslington, York YO10 5DG, UK

Abstract. It can be argued that there are robust and well-defined digital preservation strategies to deal with migrating simple digital objects such as single files (but of course the counter argument is that there are many 'simple' files that are hard to identify, contain embedded objects, or are very large). The question is: do these strategies extend easily for complex objects in general, and for visualisations and simulations in particular? This chapter is based on breakout sessions led by Dr William Kilbride on the topic.

Introduction

The participants for each session were drawn from a wide range of professional contexts concerned with digital preservation (DP) including: scientific and cultural heritage projects; private consultancies; digital assets management including creating digital preservation policy for, *inter alia*, 3D models and architectural drawings; digital repository management; DP research and development; development of research-based open source tools; digital tool preservation; long-term digital preservation; emulation research; historical research; digital humanities research; and digital imaging and digital media technologies. With this eclectic mix of contributors from a variety of user / stakeholder communities, it was important not to make assumptions about prior DP knowledge as some participants may be very knowledgeable about DP in general but be unaware of visualisation / simulation initiatives, and vice versa. With this caveat in mind, William Kilbride reviewed the main categories of challenges regarding visualisations and simulations in the DP domain, and suggested initial key responses. The groups then discussed their ideas.

Main DP Challenges and Key Responses

The first challenge to be faced is that the issue will not go away, nor will it resolve itself. A salutary note on this subject is delineated by Gartner's Hype Cycle on the introduction of new technology that charts levels of optimism over time. The cycle starts on the trigger point of 'not my problem'; moves on through 'how hard can it be' to the Peak of Inflated Expectations; descends into misery to the Trough of Disillusion when the scale of the task becomes apparent; then advances onto the Slope of Enlightenment followed by the Plateau of Productivity when realistic progress is eventually realised. To avoid this roller coaster experience, it is vital that DP plans be set out at the start of the project.

In general, digital objects such as software, simulations, visualisations, documents etc. have value (however such 'value' is calculated), and they create opportunities. However, access to such objects depends on software, hardware and human intervention, and these are liable to change over time, thus resulting in technology creating barriers to reuse. Therefore it is vital to manage data in the long term to protect digital assets and create opportunities for their reuse. In particular, DP is not just about isolated topics such as 'data', 'access' and 'risk': rather it is paramount that DP is understood to be about outcomes, especially regarding *people and opportunity*. In a nutshell, DP should be about healthier, wealthier, safer, smarter, greener people and communities.

This is not always the case, however, as DP typically makes bleak reading. To take archaeology as an exemplar, it is the case that through excavation it destroys that which it studies, so DP is crucial for this field. The technical challenges in this domain are formidable, digital achievements and take-up to date are wholly inadequate and data loss is ubiquitous. For example, the Archaeology Data Service (ADS) Strategies for Digital Data reported that the archaeological record could be decaying faster in its digital form than it ever did in the ground (Condron, Richards, Robinson, & Wise, 1999).

Challenge 1. Access and long-term use of digital content both depend on the configuration of hardware, software, the capacity of the operator and documentation.

Key Responses

1. Migration (changing the file format to ensure the information content can be read) is the most quoted, and most widely used solution. It is typically good for large-quantities of data that are well understood and self-contained (with few or no dependencies), with there being a relatively small number of formats involved.

2. Emulation (intervening in the operating system to ensure that old software can function and information content can be read). This can be used in tandem with migration; in fact migration and emulation both often require, for their realisation, deployment of elements of each other. *A vital step forward in the preservation debate is to embrace hybrid strategies deploying both migration and emulation, instead of seeing these as rival options (Anderson et al., 2010, pp. 10-18).*

3. Hardware Preservation (maintaining access to data and processes by maintaining the physical computing environment including hardware and peripherals). This is less fashionable, more expensive, but effective. It is often claimed that emulation obviates the need for preserving hardware, but this is not the case. In order to ensure that an emulation configuration is authentic, it is necessary to preserve hardware to set up benchmarks. This is particularly important for preserving the actual user experience of using the hardware and software. Computing history museums are key to this task (Anderson et al., 2010, p. 10), and the extent to which they can provide software, hardware, emulators and documentation is the subject of a scoping study in KEEP.

4. Exhumation (maintaining access to an execution environment or software services to that processes can be re-run with new data) also involves emulation and migration elements.

Challenge 2. Technology continues to change, creating the conditions for obsolescence.

Key responses

DPC Technology Watch reports / services give advance notice of obsolescence. Migration and emulation reduce the impact of changes in technology. File format registries such as PRONOM, UDFR, P2 also contribute by profiling file formats and their preservation status.

Challenge 3. Storage media have a short life and storage devices are subject to obsolescence.

Key responses

Storage media can be refreshed and in some cases can self-check, and storage densities continue to improve, offering greater capacity at reduced cost. It must be emphasized, however, that storage may only be a minor part of the solution, but not the whole solution. (In fact transferring bits from one medium to another is not necessarily inherently difficult – the problematic issues are often concerned with the *quantity* of information involved.)

Challenge 4. Digital preservation systems are subject to the same obsolescence as the objects they safeguard.

Key responses

Systems can be modular and conform to standards, and fitness for purpose can be monitored over time.

Challenge 5. Digital resources can be altered, corrupted or deleted without obvious detection.

Key responses

Digital signatures and wrappers are available that can safeguard authenticity. Also, security measures can control access to the digital material (although it must be admitted that digital preservation has not yet sufficiently well confronted security issues such as cyber attacks). It is also a real advantage that for digital as opposed to physical material, copies are perfect replicas with no degradation.

Challenge 6. Digital resources are intolerant of gaps in preservation

Key responses

Ongoing risk management can provide vital monitoring to help deal with this problem, and there are significant economies of scale to be made. It is critical that we work with colleagues in computer science towards the end that processes such as metadata creation and harvesting, and media transfer where possible, be automated. We must also be aware that data are rapidly growing in scale, complexity and appetite / importance, that is to say: the expectations we have of data.

Challenge 7. We have limited experience.

Key responses

The rapid churn in technology accelerates our research, which has been transformed over the last decade. As noted in challenge 6, this is a *shared problem*: those at the forefront of DP research such memory organisations need to be willing to appropriate solutions developed in other domains such as computer science (digital forensics, software lifecycle development) (Gladney, 2008, pp. 14, 25).

Challenge 8. DP has to cater for widely varying types of collection or interaction: so different strategies are required for different types of collection or interaction.

Key responses

Where possible we need to develop strategies that cater for discrete categories of material, such as:

- simple v. complex
- large v. numerous
- gallery v. laboratory

A helpful way of confronting the issues is to consider that DP involves three key components: Technology, Organization and Resources. With this in mind, the following list indicates the pressing questions about complex digital objects in general and visualisations and simulations in particular currently facing the DP community:

1. Are the issues data size or data complexity, or both?
2. Which is the best preservation strategy: emulation, migration, hybrid or neither?
3. Is it easier to recreate data than to secure it?
4. What does success look like and how will we recognise it (i.e. via which metrics do we use)?
5. Will the material ever be used? How do we balance delivery and accretion?
6. Does the material fit to its original mission, and whose problem is this?
7. How do we find the necessary resources and expertise?
8. What is next on the horizon: more scale and more complexity?
9. DP tools: who is making these and what are the dependencies between them?
10. Is visualisation a special case for DP, or a special DP community?

Key Issues in Response to the DP challenges

Creating a Proactive Data Management Environment

Archives require active management, a fact that many institutions have not yet come to grips with. Indeed most of the technological problems are soluble, but finding institutional resources / contexts for actively managing DP is much more intractable. In practice, many organisations – such as memory institutions – are not equipped to carry out active DP management, whether hindered by legal limits, or by lack of resources (for example cash-poor cultural institutions cannot afford the requisite level of technical expertise). Where resources are the problem, ideally these institutions need to change their economic model to secure the technical services they need. This requires a change in mindset within the institutions, and in turn means the decision makers need to have a sufficient understanding of the issues, and the data management processes involved. It is crucial, therefore, that managers understand, and are enabled to understand, the 'value' inherent in their digital data. ***Decision makers need a way of calculating the value, to an organisation, of their digital data.*** This can, in turn, lead to better decision making at an organisational / project initiation level with tangible practical outcomes, for example defining a more limited number of supported file types.

What Should We Preserve?

Historically, qualified people determined what data to preserve and what not to preserve: a process that in turn added to the value of the preserved data. In the digital domain, how do we determine the scope / nature of what is to be preserved, especially when the technical expertise needed to evaluate visualisations and simulations may be well beyond that of the curatorial staff, and relevant documentation may be too voluminous and therefore prohibitive to compile?

Digital art poses particular challenges. Many different contributors raise issues about ownership, while interactive / online art, which may dynamically change over time, and through various versions, poses questions about what actually constitutes the 'object' to be preserved. In the case of a theatre performance, it is the documentation that is preserved, rather than the performance itself. But art objects, and their various iterations, represent a somewhat different challenge. The key point is that the decision of what to preserve, and what not to preserve, ought not to be left to chance or (by default) to obsolescence, but should be a consciously-deliberated strategic decision. A useful starting point is to create an inventory of what the holdings are, and systematically to prioritise objects, or categories of objects for preservation. The Tate, for example, has started by asking key stakeholders (conservation departments, curators, artists, etc.) about the financial and legacy / heritage value of their holdings. Collection holders always have had to make archival / preservation decisions on behalf of end-users, but we could also ask to what extent end-users (in the Tate's case, website users) might also be consulted? It is important to note that digital artists may have different priorities compared with preservationists, and so it is essential not to try to shoehorn bespoke artists' definitions into existing preservation standards, but rather to expand the latter sensitively according to this particular domain.

One of the challenges is how quickly decisions about preservation / destruction of resources need to be taken in the digital age: paper-based records could be warehoused for decades before decisions were made, whereas the reliability of hard disk drives (HDDs) or USB drives is measured in a very few years only. This lack of temporal distance / perspective makes it very challenging to reliably determine what posterity will consider important. In particular, records of *processes* may not be rated highly at the moment, but may be viewed as crucial in the future. One option may be to preserve analogue versions of digital resources (e.g. print outs onto acid-free paper, which indeed is done in some organisations). This becomes particularly challenging in the case of complex digital objects, but may bear further exploration in the future.

Whose Responsibility is Digital Preservation?

This issue is not clear – is it down to the creators of simulations and visualisations, or archivists, or funding councils, or a mixture thereof? A suitable process for ascertaining such responsibility needs to be devised, otherwise the community is left with an unfunded mandate to preserve material but with no way of systematically carrying this out. Indeed, sound institutional and administrative processes are the key to making progress in this area. Mitigating against such advances is the fact that crucial technical skills are poorly distributed, making it hard for the cultural sector to recruit staff and obtain services. In short, it turns out that there are too many projects and too few 'services'. Added to that is the enormous difficulty in articulating the technical, practical and logistical requirements for simulation and visualisation DP. Preservation from the outset would be a great asset, and this leads to the need for data creation tools for preservation-ready objects.

Another crying need is for greater documentation awareness in the communities. Currently data are created by people who do not have a document culture. It is critical to know before starting a project, what you need to document in order to preserve the material created. The visualisation and simulation communities could benefit from acquiring good engineering practice where hard lessons have been learnt following on from data loss by e.g. NASA. To this end, the durability of digital objects, both complex and simple, should be inbuilt, and for this we can collaborate with and learn from the computer science / software engineering domains. In this respect, organisations such as the OPF and the DPC can play a vital dissemination role to IT students in universities by encouraging them to think about digital object sustainability. In tandem, the visualisation and simulation communities should provide guidelines / standards and templates, following the example of say the SIARD standard for preserving databases.[1]

The Problem of Preserving 'Everything'

In practice this means that the sector is not making strategic decisions to keep – or delete – objects. It is essential to define what is the boundary of a work, and then to go on to decide what should be retained or prioritised. Preservationists should be able to carry out confident deletion that would allow them to map resources against stated priorities for

[1] http://www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=en

keeping material**.** To this end, a vigorous scholarly debate about what is important would be helpful in establishing what should be kept**.** A necessary prerequisite to achieve such a goal is to consider how to engage end-users in any discussion about deletion, especially where it impinges on long-term use. In practice, research is sometimes based on other people's perceived "rubbish": for example discarded maps of Pompeii. It may be hard to gauge the importance of such ephemera – a photo or map – to a researcher building a 3D visualisation of Pompeii it could be really important. In an analogue world, we may feel sure that we know what future users would need. For digital documents, we do not know what future users will want to do, and how? There has been a paradigm shift in culture away from the physical: it seems that seeing a photo is as good as viewing the Mona Lisa (La Gioconda).

Also, it is expedient to see what can be learnt about digital preservation priorities by examining the various approaches to digitisation. In terms of time scales, it is imperative to decide a suitable DP strategy early on, not later, as time constraints are then very narrow. It is also important to be flexible, and propose different solutions for different scales of preservation, whatever type of complexity these scales encompass. Unfortunately, in most cases, urgency leads to the wrong decisions being made.

The Problem of Scale

The issue of scale is indisputably a major problem. There are six complex visualisations, so characterization is a problem of scale. We need a top-down solution, so that characterization can be undertaken systematically. The problem of scale may include not only the problem of preservation, but also the problem of access. Individual researchers in 3D visualisations may save everything, and this can result in an eclectic mess of floppies, commodore files, tar files etc., with each person thinking they are doing things properly. Whilst some standardisation is necessary, it is evident that one system will not fit everyone – it is necessary to ascertain what are their needs? Some archives preserve 5TB per day in 1,000 different formats.

Do the Problems really only start with Scale?

There are other problems concerning complexity that are not related to issues of scale. For example, preservationists of visualisations and simulations may not be able to identify the material they are given, as it takes considerable technical knowledge to be cognisant with the range of material covered by this domain. Where bespoke software / models are involved, customisation is hard to achieve, and although standardisation is desirable, it is also often an unreachable goal. A knowledge donor card / expert system would be most useful, where creators of such complex objects evince a desire / expectation to move on and leave their knowledge behind.

Do we Understand the Nature of the Problem that we have?

It is unlikely that we can have entirely generic approaches for tackling the preservation of all complex objects, due to the fact that addressing more than one domain is hard. As observed in the abstract above, it can be argued that all objects are to some extent complex. For example, PDFs making a call on a tiff file represents a very complex file format. If complex objects require bespoke – i.e. expensive solutions – who then is to be the arbiter of value? This again comes back to the question of who decides what to keep? Here, the designated community is a useful concept to establish the nature of the problem and hence the value of the data. For example the AHDS had context-specific rules that could not be transferred: they were only valid for the designated community.

Another salient approach involves determining the business case. In particular, the business cases for preservation will probably be different in different institutions. Two projects, CASPAR and SHAMAN looked at specific data to try to find a generic approach. However, it can be the case that even people looking for a single solution may end up with an integrated solution. For this POCOS symposium, it is important to establish what makes simulations and visualisations a different case. For these subject specialisms we need to understand digital preservation as a process 'in the present': encompassing good practice and assessment in the 'here and now' will be useful in the medium term.

What is the Complexity in the Object and what Effect does it have?

This could be embedded material, as in the new PDF format; nested objects; obscure file types, extremely large objects etc. or any combination thereof. An example of a complex digital object from archiving the web at the British Library highlights the difficulty of saving the commentary from the artist Anthony Gormley on the plinth at Trafalgar Square, together with the associated blog and video. The comments, the related structures and timings and overall experience were not preserved, so it was impossible to make sense of the commentary, blog or video.

One effect is that complex objects will cost more to preserve than their simple counterparts, as observed in the data security session. However, this does not necessarily mean that complex objects need complex governance. One suggestion is for e.g. a memory institution to just to save one screenshot of an entire, complex digital scene (e.g. of a 3D visualisation), whilst the creator should save all the digital components making up the whole digital scene. Here it is vital to recognize that complex objects have complex dependencies, thus it is advisable to carry out risk analysis and then prioritise problems, first ascertaining what are the weak links in the chain? Self-contained digital objects with inbuilt metadata etc. play a key role in this task.

Another point to bear in mind is that complexity can differ depending on how an object is installed e.g. under conditions of experimental extremism with flex glass panels and inputs from a person's sound, movement etc. In this situation it may be sensible to try to preserve the system and the experience, noting whether it functions correctly or not. It may be possible to preserve a predefined sequence of events, whilst allowing for the fact that random effects cannot be reproduced. In this case it is necessary to decide what is important. Documentation can play a key part in recreating a given technical environment necessary to recreate a particular experience. The data models behind the TOTEM database in KEEP / OPF[1] seek to provide metadata to allow robust descriptions of requisite technical environments.

To conclude, there is the impression that complex objects may be categorised as just being too hard to deal with, so preservationists may be tempted try to just archive them to avoid dealing with the problems. Given that such as object has no clear migration path, then the next best thing is to do a best efforts path. In the end it is a case of preserving size, environment and experience.

References

Anderson, D., Delve, J., Pinchbeck, D., Konstantelos, L., Lange, A., & Bergmeyer, W. (2010). *KEEP Project: Final document analyzing and summarizing metadata standards and issues across Europe*: EC reporto. Document Number)

Condron, F., Richards, J., Robinson, D., & Wise, A. (1999). *Strategies for Digital Data Findings and recommendations from Digital Data in Archaeology: A Survey of User Needs*. York: Archaeology Data Service (ADS), University of Yorko. Document Number)

Gladney, H. M. (2008). Durable Digital Objects Rather Than Digital Preservation [Electronic Version]. *ErpaePrints*. Retrieved 16th July 2009, from http://eprints.erpanet.org/146/01/Durable.pdf

[1] http://www.keep-totem.co.uk/

Issues of Information Security Applicable to the Preservation of Digital Objects

Andrew Ball, Clive Billenness[1]

[1] The British Library, Boston Spa, Wetherby, West Yorkshire, LS23 7BQ, UK

Abstract. This paper considers how the best practice principles of Information Security, as embodied in the still-evolving ISO 27000 series of standards on Information Security might be applicable to issues relating to digital preservation with particular reference to Complex Digital Objects. It also identifies some of the risks which arise from the application and non-application of those standards.

Introduction

There are many branches of information technology and management, but some of the techniques for management and security apply across many different discipline areas.

Because the IT industry is forward looking, it does not have a reputation for looking back. The industry always tends to focus on 'the next thing': bigger, faster and with increased capability. This gives rise to recurrent issues with backwards-compatibility and so-called legacy applications; where it is necessary to either use old data with new systems or integrate old and new information systems.

What it tends not to concern itself with is what went before. Backwards compatibility is often not taken into consideration – information security tends to concern itself with the current information life-cycle. Once data is archived or off-line, information security can start to fall away. This presents a number of challenges for digital preservation. Once data is placed in archive systems, the controls can be even less effective and the risk is not always effectively assessed.

One example of a digital preservation issue relates to old electronic musical instruments – many of the older synthesisers from the 1970s and 1980s had unique sounds which still have a place in modern music. However, many almost became extinct as newer instruments were produced. The sounds made by these have been rescued by a number of projects and remain available to a new, modern audience. These were almost lost but were digitally re-enabled.

Organisations concern themselves with 'governance', but this is often a euphemism for 'compliance'. The problem can arise where those who design regulations are not sufficiently close to the actual business and so design controls that create unnecessary barriers to normal operations. Such regulation (combined with business targets) tends to create perverse incentives for operational staff to circumvent the procedures rather than comply with them in order to 'get the job done'.

"Tick-box" governance can therefore be a serious problem for a business, since there is no sound basis for the compliance framework that has been adopted. It is not a substitute for 'doing it properly'. The key question should be "with what are we complying?" And it is important to select the correct standard to adopt and understand its scope and application.

There will always be tensions between end-users, IT operations and information security specialists. There are many situations when rules and restrictions are imposed on the use of information systems without any attempt to explain the reason for these restrictions. What is needed is dialogue between the different interests to create a consensual approach that will encourage informed compliance. Humans are at their most creative when seeking ways to bend rules. For this reason, rules that make sense and are attuned to the business purposes are more likely to be complied with.

Information Security is built on three principles:

- Confidentiality
- Integrity
- Availability

To achieve good information security, all three principles must be addressed.

In most organisations, Confidentiality is well understood and complied with; legislation and regulation imposes duties and responsibilities on organisations where appropriate.

Integrity is similarly well understood, particularly as it relates to data quality. Problems can arise, however, where there is a disconnect between the people who populate systems with information and the end-users of that information. One example of this is in medical record systems where clinicians record information about examinations and treatments and focus only on the medical aspect of that data capture; they may potentially record the incorrect activity code as there may be several covering their specialty. Later on, these codes are used by finance departments with an entirely different interest in the data – i.e. determining the costs of treatment. This can lead to difficulties in determining and applying the correct charges.

In the National Health Service, this can lead to hospital and primary care budgets being adversely affected through internal mischarging. This can, in turn, lead to reductions in quantities of treatment provided as budgets appear to have been fully consumed.

Thus, making the connection between information providers and end-users is critically important, as well as understanding the impact on an organisation of errors occurring in data. This applies not simply within the field of knowledge of the information provider but also by other, future users.

The least well-understood aspect is undoubtedly availability. This is partly because there is an inevitable tension between the principles of Confidentiality and Integrity and that of Availability. In short, the more effective the controls to secure data and protect it, the less accessible it becomes. In addition, when data is archived, the question of long-term continuing accessibility is not always considered.

In considering “what makes good information”, there are three further principles which were proposed in a 2008 report by the Audit Commission[1]:

- Relevance
- Presentation
- Quality

Relevance requires that the information being considered is applicable and appropriate to the business overall or the decision being taken.

Quality requires that the information be of appropriate quality – i.e. ‘good enough’ for the purpose to which it will be used and at the time at which it will be used. This recognises that there will be times when information will be approximate or incomplete but provided these deficiencies are recognised, this is an adequate control.

[1] The Audit Commission (2008) In the know. ISBN: 1-86240-545-X. Retrieved from http://www.audit-commission.gov.uk/nationalstudies/localgov/Pages/intheknow.aspx

Finally, information must be presented in a clear, concise and understandable way, This means that sometimes information must be summarised appropriately to the level of the reader.

The Information Security Community has evolved an extensive series of standards to codify some of these issues. Many standards exist which cover Data Protection, Business Continuity and Information Security. An example of a standard is the ISO27000 series of standards on aspects of Information Security (ISO27001 – ISO27006)[1] One standard in the ISO27000 series – IS027037 – "Evidence Acquisition Procedure for Digital Forensics" has potential relevance to digital preservation activities as its summary explains:

> *"This International Standard provides detailed guidance that describes the process for recognition and identification, collection and/or acquisition and preservation of digital data which may contain information of potential evidential value."*

It must be noted however that the focus of this standard is digital preservation and the law, so its applicability may be limited. The current target publication date for this standard is 26/10/2012.

Standards are attractive to organisations that wish to obtain and demonstrate certification of compliance. However, it is important to know the scope of any certification, and also to be aware of how recently any certification process was undertaken. Compliance and certification can sometimes be quite specific and limited in scope within the organisation overall.

ISO/IEC 27002:2005 (Information technology - Security techniques - Code of practice for information security management) contains a number of sections that are relevant to digital preservation specialists. These are considered below.

Security Policy

It is dangerous to 'clone' one security policy from one organisation to another without engaging with local users to identify specific individual needs. There is no guarantee that a policy that will work effectively for one organisation will be equally effective if simply imposed on another. While consistency in the creation of policies is laudable, and the use of a framework tends to promote effective outcomes, it is important not to be constrained by this approach and so fail to reflect the operational needs of the organisation. There is no substitute to working closely with end-users to identify what will be effective within the specific organisational context and where some compromise might be necessary in the interests of operational efficiency and improved compliance

Security Organisation

It is important to engage with both stakeholders and senior managers on matters of information security. It is interesting to note that a number of high-profile losses of media carrying sensitive data in both the public and private sectors have helped to raise the

[1] http://www.27000.org/

profile of information security and governance within organisations. Few organisations are unique, so it can be very helpful to seek analogues with which to compare one's own organisation's approach in order to learn and promote mutual improvement.

Third-party suppliers are another important component of an organisation's security arrangements. It is vital to ensure that there is a good mutual understanding of one another's information security arrangements to ensure that, where necessary, ongoing data retrieval can be supported following mergers, discontinuation of contract or even one or the other party going into administration.

Information Security arrangements must also include a 5- or even 10 to 20-year perspective if they are to adequately serve digital preservation specialists. There are sadly only a very few cases where security plans look this far ahead.

Asset Management

This considers how digital assets are 'owned', classified and managed. Where the classification of a digital object is intended to restrict access to it, definitions of that classification may change over time. If an object is archived while classified with a restriction on access, does an adequate procedure exist to review or amend the classification at a later date. In the case of a Complex Object where only a part of it is subject to a restriction, immediate questions arise as to how and where to store different components of the object. This process will need to take account of how links between the different components are to be maintained.

In addition, a procedure must be created to allow for future de- or re-classification where necessary. This may not be well-recognised at the point of archival, leading to difficulties in retrieval in the distant future.

In order to create effective processes, arrangements for both 'ownership' and also 'succession' over a long period of time must be established and documented, covering the context and meaning of any classifications that have been applied.

The widespread use of Digital Rights Management to protect digital intellectual property – e.g. games – can also present a challenge where an object is retrieved and the holder of the Digital Rights is not easily identifiable in order to obtain permission for use or to re-access the contents of an e-book. The intent of an archivist or another later user of preserved digital objects as opposed to an original owner must also be taken into account. An original owner will want to use digital content for its original intended purpose and so DRM restrictions are appropriate; but the intended usage by an archivist may differ to the extent that the DRM is no longer appropriate.

People Security

All processes relating to personnel management (hiring, induction, leavers etc.) have a strong link to information security. It is a widely-quoted saying in information security that 'people are the biggest risk'. One of the issues within this area is that of 'tacit knowledge' – the operationally important, but undocumented, information that staff acquire over a period of employment. Especially when considering long-term digital preservation issues where this 'tacit knowledge' may no longer be available later on when a future user of a digital object attempts to withdraw it from archive. There are many reasons why information is not shared. Although some people do undoubtedly deliberately retain knowledge in order to maintain or increase personal influence, in many cases, organisations do not provide an adequate mechanism for such sharing to occur. It is therefore important to ensure that mechanisms exist to ensure that this information can be gathered and organised to enable future access to archived data.

Physical and Environmental Security

The question which must be answered is that of who will be responsible for the physical security of the equipment on which digital objects will be stored and processed. With increasingly complex physical environments, often dependent on 'virtual' machines or assets stored in the Internet Cloud and potentially 3rd-party managed, it should not be automatically assumed that the service provider is heavily focused on issues of information security. Instead, providers should be asked to demonstrate any certifications and (as referred to above) the scope and date of those certifications. This should present no problem to competent and experienced organisations.

When selecting a supplier, it is also advisable to ensure that they operate a continuous risk assessment approach and respond proactively to security threats and incidents to ensure that they retain effective defences against evolving security risks.

When considering access to working hardware, digital processing equipment rapidly becomes obsolete. While enthusiasts still have the ability to source such equipment via online auction sites and used equipment dealers, this would be a high-risk strategy for an organisation with a heavy investment in digital objects.

There are two options for addressing hardware obsolescence. The first is to maintain a 'hardware library' complete with all necessary associated software on which the original environment of a digital object can be retained. The second is to create an emulation environment on which the original environment can be replicated. It must be borne in mind that most, if not all, hardware environments will be at least partially dependent on the use of Programmable Read Only Memory chips (PROMs). In addition therefore to the risk of simple electro-mechanical failure within a hardware device, there is also a lifespan to PROMs which is currently estimated to be anywhere between 10 and 30 years. Such a hardware library can therefore only provide a time-limited solution with an ever-increasing risk of both hardware failure and partial or complete media failure.

The alternative is to establish an emulation framework to enable the original hardware/software configuration to be replicated in software within a modern environment. This can then continue to evolve with each subsequent generation of computer system. Recent European Projects such as Planets and KEEP have undertaken substantial research in this area – work that continues to be developed by the Open Planets Foundation (OPF).

Communications and Operations

Security should be considered as part of the operational lifecycle. While security considerations are now frequently considered at the original specification stage of a procurement of new equipment or systems, it very likely that archiving and long-term preservation issues will only be considered shortly before de-commissioning. Archiving and preservation should be designed into the system and into operations from the outset

Access Controls

There is generally a heavy reliance on passwords or variants of password systems such as pass phrases or electronic tokens. Digital Rights Management is increasingly being used to protect commercially-distributed creative material. When preserving digital objects, it is also necessary to document and store the authentication. It is poor practice to store authentication keys in proximity to the object to be protected, Instead, there should be a separation of tokens and material to be protected. This, however, requires more complex arrangements that are sustainable over an extended period of time.

In organisations, employees tend to accrue access rights as their career progresses, never losing those applicable to their previous post but continuing to acquire those necessary for their new role as there are inadequate arrangements for revocation of legacy permissions. This is comparable to giving a front door key to every tradesperson who ever calls at a house and never retrieving them or changing the lock. No individual would find this to be an acceptable practice in their own home, yet this kind of behaviour is all too common within organisations that lack effective processes to revoke access rights.

The same issue arises when dealing with the access rights for archived digital objects. When dealing with complex digital objects, the issue is magnified, as all access rights must remain synchronised. Failure to revoke access rights to just one data stream creates a risk to the integrity of the entire object if a user who retains legacy access rights but is unaware of some new relationship between that data stream and other, related objects, makes an alteration which compromises that relationship.

Information Systems Security

Information Systems Security should be considered in terms of its relationship with all aspects of the information systems lifecycle from specification, through project management to implementation, operation, change control and finally decommissioning.

Security risk assessments should consider long-term preservation issues, but rarely do so. Typically, Privacy Impact Assessments, which are conducted on new systems, focus on Confidentiality and Integrity but less so on Availability even though this is an equally important aspect of security. The earlier that long-term preservation is considered within any project, the less likely it is that a solution will be implemented which is incapable of meeting the long-term preservation needs of the organisation. It is perfectly conceivable that a systems solution that would otherwise meet all requirements might be rejected at the feasibility stage if it fails to meet preservation requirements and so undermines the overall business case for the project.

Security Incident Management

Organisations should learn from incidents that will occur in any organisation, and should use the lessons learned to improve their security arrangements. What may not occur, however, is that any revisions to security arrangements are also applied to digital objects in long-term preservation. Too often, revised arrangements are applied to operational systems. It is notable that hackers frequently target backup systems because access controls are not usually as well-implemented. Thus systems used for long-term preservation, with weaker controls, might provide an easy means of penetrating the newer, perhaps more sensitive, live operational environment. By definition, a backup system contains all the data held in the live system, and so merits equal security.

Business Continuity Management (BCM)

This is focused on risk assessment, with two aspects to be considered.

Firstly, there is Disaster Recovery: restoring the technical environment. Secondly there is Business Continuity – maintaining the day-to-day activities and services of the organisation until the technical environment has been restored. Business Continuity requires that documentation and key operational information be available even if the main buildings are inaccessible or otherwise compromised – for example by fire or flood. Loss of a main site might result in the loss of all the configuration information necessary to restore services, even if an alternative site was available.

For digital preservation, there are some special long-term risks to be considered. One of these is Climate Change. Long-term preservation activities demand a long-term approach to risk assessment. Forecasts of the levels of future flood-plains and impacts of rising sea levels on coastal areas over a 10-year timeline form an important part of planning activities when identifying the locations of long-term archiving and preservation sites. This level of planning should be applied.

Compliance

Compliance focuses on people. It is centred on the cultures within an organisation and addresses legislative requirements and/or organisational policies. In not all organisations, however, is there sufficient effort invested in communicating policies to staff and in monitoring the extent to which policies are effective. Experience shows that people will often commit to and speak in support of security policies and yet their observed behaviours are entirely in contravention of those policies, often through lack of appreciation of the potential risks to the integrity of systems and data of non-compliance.

The key method of combating these vulnerabilities is through education. Although many organisations rely on sanctions and disciplinary action to enforce policies, incentives and engagement are generally more effective.

When addressing preservation issues, organisations should test their plans for retrieval of archived information. Without tests, there is no assurance of success of any retrieval plan.

Beyond ISO27002, modern information tools have created new opportunities and new threats to the integrity of an organisation's data and its long-term preservation. Web 2.0 technologies have led to a rapid growth in user-generated content, and archiving the content of these sites is a complicated process and even major service providers experience difficulties in performing this task. Modern document management tools empower users to store information in a structure of their own choosing. Where complex objects are created by interconnecting internal and external data sources, maintaining these connections when archiving data can become a very intricate operation. This means that the documentation of object structures and relationships becomes an essential requirement if accessibility is to be maintained.

Similarly, e-mail systems can have a wide variety of formats of digital objects embedded within individual messages. In the event that an organisation changes its e-mail service, it can be very difficult to retrieve the contents of individual e-mails or to identify and retrieve a chain of related e-mails. Many organisations do not even retain licences or software for discontinued systems of this type, therefore diminishing the probability that the contents of these systems can ever be recovered and accessed in the future.

Organisations have rapidly become very reliant on large-scale but transient brands of web-based software without any real consideration of how the corporate data stored by them can be backed up and retrieved in the event of the service provider ceasing operations. The rapid growth of hosted 'blogs' (estimated at 10,000 new blogs per day) is one example of how information is gathered very quickly but is very hard to manage and preserve. These modern brands are very fluid, and there are regular announcements of organisations merging or dividing into separate entities, with the attendant difficulties in maintaining links between objects. Overall, the pace of technological redundancy is accelerating, and this presents new difficulties for organisations concerned with digital preservation.

Summary

Availability is the key principle for digital preservation practitioners, although Integrity and Confidentiality are no less important. There is a risk, however, that when organisations consider information security, there can be an over-emphasis on Confidentiality and Integrity to the detriment of Availability. As a control environment is defined, those concerned with long-term digital preservation must challenge the balance of focus applied to each of these three principles to ensure that their business needs are adequately catered for.

There is a tendency in projects to treat the costs of preserving legacy data as a part of the project to replace the existing system. This is an error, as, if there is a requirement to ensure the long-term preservation of digital objects, these are actually 'business-as-usual' costs.

Digital Preservation requirements should be taken into account as early as possible in the life-cycle of a system, if possible during initial specification, and not left until these become an urgent issue shortly before, or even at the point of, system decommissioning. Digital Preservation activities could represent a significant element of cost within the overall Business Case for a system and it can be expensive to overlook it or not take into account during assessment of different options.

Risk assessments must not cease at the end of the operational life of systems and data. They must be continued onward into the preservation space as part of business-as-usual, with controls being updated in parallel with live data.

References

The Audit Commission (2008) In the know. ISBN: 1-86240-545-X. Retrieved from http://www.audit-commission.gov.uk/nationalstudies/localgov/Pages/intheknow.aspx

The Impact of European Copyright Legislation on Digital Preservation Activity: Lessons learned from Legal Studies commissioned by the KEEP project

David Anderson

CiTech Research Centre Director and Co-Leader Future Proof Computing Group, School of Creative Technologies, Eldon Building, University of Portsmouth, PO1 2DJ, UK

Abstract. Digital preservation activity in the European Union takes place within a complicated and often contradictory legislative landscape. Over and above national law, stands the European Community framework – which, although meant to be incorporated into member state legislation, is not uniformly or completely implemented across the whole of the EU. Finally, certain non-EU legislation, as well as international understandings and treaty obligations such as the Paris Convention for the Protection of Industrial Property (1883), and the Berne Convention for the Protection of Literary and Artistic Works (1886), all play their part in determining the precise legal status of preservation actions. During the first year of the KEEP project two legal studies were commissioned[1] to explore the impact of European law on the project's proposed programme of work. The purpose of the present document is to articulate, as far as is possible, the main conclusions of the KEEP legal studies in layman's terms. Naturally, this involves some diminution of legal rigour in consequence of which the present document should not be regarded as legally definitive.

Introduction

The KEEP project is the first EC-funded project to concern itself primarily with an emulation-based approach to digital preservation. From the outset it was recognised that there was some potential for the project to encounter unique legal issues. Emulation involves the creation of software that permits one hardware platform (computer) to 'mimic' the behaviour of an entirely different hardware platform. In a preservation context this has enormous potential as it offers the possibility to run software originally designed for a now obsolete computer on the latest hardware although there are, of course, considerable technical challenges which need to be addressed before this potential can be achieved. There are also obvious copyright issues which need to be considered when writing or using software that attempts to reproduce exactly the behavioural characteristics of third party code.

An emulation-based approach to preservation has the advantage of avoiding any need to make alterations to preserved files in order to make them accessible on modern machines. In this respect emulation is very different in character from 'migration'-based preservation techniques that 'convert' old file formats into forms which run on new platforms. The primary problem which the KEEP legal studies sought to address concerns 'media transfer', which is the process of moving computer files from their

[1] The legal work was sub-contracted to Bird & Bird and was presented to the KEEP consortium in the form of detailed reports.

original storage medium (5.25 in floppy, magnetic tape, etc.) onto a managed storage system within a library context. This is essentially a process of copying software.

The KEEP project initially proposed to create a framework which would agglomerate a number of distinct media transfer tools, with the aim of creating a "one-stop shop" for media transfer within a digital preservation context. It was recognised from the outset that transferring copyrighted material from one medium to another was subject to legal regulation both with respect to what might legitimately be transferred, and what use might permissibly be made of material after transfer had taken place. One aim of the legal studies was to delineate more precisely the legal boundaries of such a framework.

Rights holders very frequently look to protect their rights over digital material by means of encryption, password protection or by other so-called "Technical Measures of Protection" (TMP). There are a number of situations in which a library or archive may have a need, or even a legal responsibility, to bypass TMP but cannot agree with the rights holder a means by which this might be achieved. On some occasions the rights holder cannot be identified or has ceased trading or may be unwilling to cooperate. The legal studies therefore sought to clarify the state of the law with respect to bypassing TMP.

The KEEP consortium contains three national libraries each of which has a 'legal deposit' responsibility within its own legislative framework. Unsurprisingly therefore one of the topics which our legal advisors were asked to explore was the degree to which "legal deposit' is recognised at the Community level and how far exercising legal deposit responsibilities might attenuate the operation of the law on copyright.

The stakeholders in the KEEP project were particularly interested in so-called 'complex digital objects', such as multimedia works[1] (e.g. computer games), and interactive educational software, and sought greater clarity about whether the complexity of digital objects has any consequences in law. Furthermore, the KEEP project sought legal clarification on any restrictions that might apply to the use of computer files following media transfer. Finally, the project sought clarification about any limitations that the law might place on batch (i.e., large scale) transfer of software, since the national libraries likely to be involved in this activity have vast quantities of software titles that need to be preserved.

The KEEP legal studies threw up two important issues: making copies of digital materials (media transfer) and making these copies available to users. Libraries and archives have no general right of reproduction but may reproduce (or transfer) digital material only in certain specified cases. The exemptions that libraries enjoy are sufficient to permit at least some of the activities necessary for preservation. However the inconsistency between national and Community laws and the lack of clarity on key terms (e.g., multimedia works) give rise to confusion at the margins. There is a tendency for national legislation to be both more permissive than Community law, and for it to provide a greater degree of detailed governance. Unfortunately this leads both to inconsistency between member states, and to national regulation that is, in certain key areas, almost certainly incompatible with Community law.

Complex though the legal landscape is, a number of consistent and clear messages have emerged from the investigations carried out both into EC law and the three national jurisdictions.

[1] For the purposes of the legal studies study, the term 'multimedia works' is understood as combining audiovisual, software and, as the case may be, database elements along with off-the-shelf software programs and databases considered on a standalone basis.

European Copyright Law: A Very Brief Overview

Copyright laws have generally attempted to balance ensuring a reward for creativity and investment, and the dissemination of knowledge. The preamble to the Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, makes clear that the overriding purpose of harmonising copyright regulation within the EC is to ensure that competition in the internal market is not distorted. This is fully in line both with the Treaty establishing the European Community, and with the general notion that the primary purpose of copyright law is to promote knowledge by establishing, for authors and creators, a temporary monopoly over their output, thereby permitting them to protect and stimulate the development and marketing of new products and services, and the creation and exploitation of their creative content.

The rights granted to authors while very extensive are not unrestricted, since there is recognition that creating an absolute monopoly would have the effect of stifling rather than promoting markets. Where the rights of authors are seen as conflicting with a public interest, various exemptions are provided. However, the onus is on those who wish to make use of exemptions to copyright protection to demonstrate clearly that they are properly entitled so to do, and that they have complied with any restrictions placed on the use of material reproduced under a copyright exemption. This is, in practice, somewhat difficult to accomplish.

While the purpose of the KEEP legal studies was to come to a view about the legal status of the work proposed to be carried out within the KEEP project, the findings have implications for any institutional progranmme of Digital Preservation. Something of the complexity that examining the legal issues involved in Digital Preservation involves may be conveyed by carrying out a ‘keyword’ search at legislation.gov.uk to see how many pieces of UK legislation touch on a given topic.

Table 1. Search results, legislation.gob.uk

Keyword	Pieces of legislation
Copyright	>200
Software	>200
Database	167
Intellectual Property Rights	163
Trademark	74

The complexity indicated by the result of this search should make us hesitant in assuming that it is possible, within the context of the KEEP legal studies, to arrive at anything other than the most tentative of conclusions.

The Community Legal Corpus

Principal Legislation at the Community Level includes:

- **The Information Society Directive** (Directive 2001/29/ EC of 22 May 2001 on the harmonization of certain aspects of copyright and related rights in the information society);

- **The Computer Programs Directive** (Directive 2009/24/EC of 23 April 2009 on the legal protection of computer programs (Codified version replacing the abrogated Directive 91/250/ EEC of 14 May 1991);
- **The Database Directive** (Directive 96/9/EC of 11 March 1996 on the legal protection of databases);
- collectively referred to as the "Community Framework".
- **Resale Rights Directive** (Directive 2001/84/EC 27 September 2001 on the resale right for the benefit of the author of an original work of art);
- **Rental Directive** (Directive 92/100/EEC/EC 19 November 1992 on rental right and lending right and on certain rights related to copyright in the field of intellectual property).

The following general rights are protected:

- **reproduction** for authors, performers, producers of phonograms and films and broadcasting organisations;
- **communication to the public** for authors, performers, producers of phonograms and films and broadcasting;
- **distribution** for authors and for performers, producers of phonograms and films and broadcasting organisations;
- **fixation** for performers and broadcasting right of rental and/or lending for authors, performers, producers of phonograms and films;
- **broadcasting** for performers, producers of phonograms and broadcasting organisations;
- **communication** to the public by satellite for authors, performers, producers of phonograms and broadcasting organisations;
- **reproduction, distribution and rental** for authors of computer programs.

The 'Three-Step Test'

Critical to understanding the general framework within which IPR legislation operates is the so-called 'three-step test', which first appeared in the Berne Convention, and has come to be regarded as a cornerstone of international copyright regulation, and imposes constraints on the possible limitations and exceptions to exclusive rights under national copyright laws.

The three-step test applies to limitations and exceptions to copyright protection and specifies that they will:

- be confined to certain special cases;
- not conflict with a normal exploitation of the work;
- not unreasonably prejudice the legitimate interests of the rights holder.

Limitations and Exceptions within Community Law: the Information Society Directive

Temporary acts of reproduction which are transient or incidental [and] an integral and essential part of a technological process and whose sole purpose is to enable: a transmission in a network between third parties by an intermediary, or a lawful use of a work or other subject-matter to be made, and which have no independent economic significance.

The directive permits Member States to make provision exceptions or limitations to the right of reproduction and/or communication in some twenty cases. Of these the following four are of direct relevance to organisations responsible for Digital Preservation:

- in respect of specific acts of reproduction made by publicly accessible libraries, educational establishments or museums, or by archives, which are not for direct or indirect economic or commercial advantage (Art. 5, 2(c));
- incidental inclusion of a work or other subject-matter in other material (Art. 5, 3(i));
- use in connection with the demonstration or repair of equipment (Art. 5, 3(l));
- use by communication or making available, for the purpose of research or private study, to individual members of the public by dedicated terminals on the premises of establishments referred to in paragraph 2(c) of works and other subject-matter not subject to purchase or licensing terms which are contained in their collections (Art. 5, 3(n)).

The Directive therefore permits limited rights for memory institutions to make copies for the purpose of preservation, but not for general communication. For reproduction to be permissible it must be permitted under national law, and should not conflict with a normal exploitation of the work, nor unreasonably prejudice the legitimate interests of the rights holder.

The Information Society Directive is typically regarded by the academic community as a victory for copyright-owning interests (publishing, film, music and major software companies) over content users' interests. The list of exceptions outlined in the Directive has achieved a certain degree of harmonization but it is important to note that Member States have no power to introduce new limitations not already included in the Directive. This has the unwelcome effect that Member States possess no independent ability to keep their legislative frameworks up to date with unforeseen technological developments.

The KEEP legal study concluded that media transfer should primarily be assessed under the Computer Programs Directive and the Database Directive.

Limitations and Exceptions within Community Law: the Computer Programs Directive

The Computer Programs Directive gives the rights holder exclusive rights to authorize:

- **the permanent or temporary reproduction of a computer program** by any means and in any form, in part or in whole; in so far as loading, displaying, running, transmission or storage of the computer program necessitate such reproduction;
- **the translation, adaptation, arrangement and any other alteration of a computer program** and the reproduction of the results thereof, without prejudice to the rights of the person who alters the program;
- **any form of distribution to the public**, including the rental, of the original computer program or of copies thereof.

However, the legal owner of a program is assumed to have a licence to:

- create any copies **necessary to use the program** and to alter the program within its **intended purpose** (e.g. for error correction);
- make a **back-up copy** for his or her **personal** use;

- decompile the program if this is necessary **to ensure its operates** with another program or device, **but not for any other purpose**.

None of the exceptions set out in the Directive expressly serves the purpose of institutional Digital Preservation and the Directive does not provide for an related to legal deposit requirements or for scientific, study or education purposes that would be similar or close to those set out by Article 5.2 (c) and 5.3 (n) of the Information Society Directive.

As a result, reproduction of computer programs carried out by institutions like libraries and museums, even when authorized under national laws, is in conflict with the Directive.

Limitations and Exceptions within Community Law: the Computer Programs Directive

The Database Directive harmonizes the treatment of databases under copyright law, and creates a new *sui generis* right for the creators of databases that do not otherwise qualify for copyright protection. A database is defined as "a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means" (Article 1).

The overall objective of the Directive is to provide:

- copyright protection for the intellectual creation involved in the selection and arrangement of materials;
- *sui generis* protection for an investment (financial and in terms of human resources, effort and energy) in the obtaining, verification or presentation of the contents of a database, whether or not these have an intrinsically innovative nature.

The Database Directive gives the rights holder the exclusive right to authorize:

- temporary or permanent reproduction by any means and in any form, in whole or in part;
- translation, adaptation, arrangement and any other alteration;
- any form of distribution to the public of the database or of copies thereof (subject to the exhaustion of rights);
- any communication, display or performance to the public;
- any reproduction, distribution, communication, display or performance to the public of a translation, adaptation, arrangement or other alteration.

Member States are allowed to provide limitations of rights in the following cases:

- reproduction for private purposes of a non-electronic database;
- for the sole purpose of illustration for teaching or scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved;
- for the purposes of public security of for the purposes of an administrative or judicial procedure.

It is reasonable to assume that a significant proportion of databases, whether made available on a standalone basis or embedded in a multimedia device, will be protected by copyright. Databases put on the market or otherwise made available to the public on tangible media generally offer more than a simple list or catalogue of items or data and are likely to be eligible for copyright protection under national laws within the EU.

None of the copyright related exceptions or *sui generis* rights offered by the Directive is relevant for the purposes of institutional Digital Preservation.

Implications of the rules on Technological Measures of Protection

Many works are made available in a form to which technical measures have been applied to prevent or restrict the use that may be made of them. This might take the form of a simple password protection scheme or may involve considerable technical sophistication.

Provisions related to technological measures and rights management information originate from the World Intellectual Property Organisation (WIPO) Copyright Treaty (WCT) and the WIPO Performances and Phonograms Treaty (WPPT). The WCT and WPPT mandate the provision of adequate and effective legal remedies against anyone knowingly performing an act that may induce, enable, facilitate or conceal an infringement of any right related to rights management information.

The Information Society Directive (2001/29/EC) recognises the "need to provide for harmonised legal protection against circumvention of effective technological measures and against provision of devices and products or services to this effect". It stipulates that "Member States shall provide adequate legal protection against the circumvention of any effective technological measures, which the person concerned carries out in the knowledge, or with reasonable grounds to know, that he or she is pursuing that objective." However it also permits Member States to be given the option of "providing for certain exceptions or limitations for cases such as educational and scientific purposes, for the benefit of public institutions such as libraries and archives".

Our investigation has shown that the potential for exemptions is quite limited and does not extend to permitting the creation or use of tools by individuals to bypass TMP. It should be noted that even in the limited situations where TMP may legitimately be circumvented, subsequent use of the transferred or copied material is extremely limited.

Other Key Legal Issues Examined

On-line Dissemination of Digital Material

In line with the general approach to the knowledge economy, libraries and others increasingly try to bring information to users rather than requiring users to come to information. Our research clearly indicates that the current legislative framework on copyright and digital material does not support this approach and contains restrictions on making digital material available on-line or off-site.

Legal Deposit Status

Community law provides no specific exceptions in respect of Legal Deposit. National legislation which grants special exemptions to legal depositories permitting them to engage in preservation activities not open to others is almost certainly inconsistent with EC law.

Multimedia works

No definition of ‘multimedia works’ exists either in Community law or in the national jurisdictions examined. There is, however, general agreement on taking a distributive, approach in which each component part of a multimedia work: audio, graphics, software, database, etc., is considered separately. Since multimedia works are not, in general, made available on computer platforms in such a way that individual elements can be removed from the whole, this means that, in practice, multimedia works enjoy the strongest protection under law that is available for any of their constituent parts. This has a significant impact on the preservation of multimedia works. Libraries (and others) are placed under a responsibility to perform a detailed assessment for each individual multimedia work they intend to preserve (or transfer from one storage medium to another) to establish the level of legal protection it enjoys. Given the scale on which national legal depositories are required to operate, such an individual assessment is impractical.

General conclusions of the KEEP legal studies

The KEEP legal studies concluded that with respect to Community Law:

- None of the exceptions set out at the Community level serves adequately the purposes of memory organisations in going about their digital preservation activity.
- Community Law does not provide for legal deposit requirements.
- Community Law does not provide for scientific, study or education purposes across the full range required for memory organisations.
- Reproduction of computer programs and databases even when carried out by memory organisations and authorized under national laws, is in conflict with Community Law.

Acknowledgements. The Keeping Emulation Environments Portable (KEEP) Project is co-financed by the European Union’s Seventh Framework Programme for research and technological development (FP7), Grant Agreement number ICT-231954.

Remember, Restructure, Reuse – Adding Value to Compound Scholarly Publications in a Digital Networked Environment

Anouar Boulal, Martin Iordanidis, Andres Quast

North Rhine-Westphalian Library Service Centre (hbz). Jülicher Str. 6. 50674 Köln , Germany

Abstract. In this paper we present a formal technical approach which structures, disseminates and reuses complex digital objects of potentially any format. The applicability to existing scholarly publications systems is paramount for the *eco4r* project in which these results were achieved. The technical approach introduced utilizes high level descriptions of complex digital objects which are used for the exchange of contextual aggregations and may be reused by digital preservation technologies. Practical outcomes of the approach introduced in this paper are plug-ins for the repository systems Fedora and OPUS and an Overlay Journal demonstrator, respectively.

Introduction

Within the last decade, the nature of scientific publications has changed increasingly from monolithic to more complex digital objects. New techniques, guidelines and frameworks have emerged, enabling researchers to enhance their publications with any kind of supplementary digital material. Thus, researchers can provide online access to their scholarly works along with any associated material and results illustrating the entire discovery process. In a best-case scenario, this leads to more comprehensible and documented publications. Scientists can provide intermediate outcomes and materials during research in order to illustrate the genesis of their final results.

Scientific publications then appear as bundles of digital objects including text, visualizations, research data, supplementary materials and any kind of other digital objects. In turn, they bring along an internal complexity of object-to-object relationships and object level metadata, which needs to be handled by digital research environments. Electronic publication systems are challenged by this increased complexity within their data sets, as they have to ensure that all parts of the publication are provided in correct structure, context and meaning. For instance, a file system randomly storing files in a directory cannot reproduce the relations between files named “data1.xsl”, “data2.xsl” and their combined results in a file called “result_data.xsl”. Thus, much of the semantic information stored on the level of data structure is being lost.

Instead, it is recommended to store data in a structured way in order to capture the components, their metadata information and the relationships between them. Ideally, any existing infrastructure in the field of scholarly publishing will be adapted and expanded for this purpose. In this article, we will use the term Compound Scholarly Publication (CSP) to describe complex forms of scientific publications, although we are aware of the existence of many other terms for similar publications (for more details see the Defintion and Scope section).

Several organizations and projects (DRIVER[1], Europeana[2], SURFfoundation[1], OAI-ORE[2]) have addressed the issue of Compound Scholarly Publications. However, the

[1] http://driver-repository.eu

[2] http://www.europeana.eu/portal

domain of complex scientific publications is still in its infancy. The current infrastructure for academic research still focuses on storage and dissemination of individual resources. This is because many repository systems do not provide functionality to create, store and manage Complex Objects. On the other hand, the creation of complex publications starting from existing data can be difficult, time-consuming and costly. Since practical implementations within existing repository infrastructures are not well established yet, we put special emphasis on the availability of software plug-ins in the eco4r[3] project.

Technical Approach

The aim of the eco4r project is to address and meet practical requirements in the field of Compound Scholarly Publications. The project examines the impact of complex shapes of publications on existing information systems, on semantic technologies such as Linked Data[4] as well as on Long Term Preservation.

The eco4r approach was started with a context analysis of existing complex objects in two exemplary scholarly repositories, Fedora[5] and OPUS[6], the latter being a wide-spread repository system at German universities. Based on these results, we have formulated guidelines for generating OAI-ORE resource maps from existing repositories (Boulal et al, 2010). At this early stage, resource maps were created manually in order to discover the requirements of data completeness as well as the technical possibilities to represent them. From the perspective of digital preservation, this first stage of analysis addresses the definition of significant properties at a structural level. Secondly, we developed a data model for Compound Scholarly Publications on the basis of OAI-ORE and the FaBiO ontology. Based on this data model, we created linked open data sets which can aggregate data that originates from several repositories and their resource maps.

In order to display aggregated content from multiple resources we recognized the demand for an integrated visualization layer. Being within the domain of scholarly publishing, we developed components for an 'Overlay Journal'[7] that dynamically creates topical data sets according to DDC classification. These components are as follows:

- OAI-ORE plug-in for generating resource maps from Fedora repositories
- OAI-ORE plug-in for generating resource maps from OPUS repositories
- OAI-PMH based harvester for OAI-ORE Resource Maps
- RDF triple store for storing the Resource Maps.

A proof of concept for the Overlay Journal will be available by the end of the eco4r project.

Definition and Scope

Several terms have been used to describe identical or very similar concepts of Compound Scholarly Publications in technical literature. The term Enhanced Publications (Place et

[1] http://www.surffoundation.nl
[2] http://www.openarchives.org/ore
[3] http://www.eco4r.org
[4] http://www.w3.org/DesignIssues/LinkedData.html
[5] http://www.fedora-commons.org
[6] http://www.opus-repository.org/index.html
[7] http://www.earlham.edu/~peters/fos/guide.htm#overlay

al., 2008) was introduced to describe publications which combine heterogeneous, but related web resources. Other researchers propose the concept of Enhanced E-Theses (Ruijgrok, Slabbertje, & Van Luijt, 2009) and continue the concept of traditional academic publications. Both terms refer to the idea of a primary publication – i.e. in terms of a full text – which is enhanced by supporting materials. Our approach is based on the concept of aggregated web resources which each may act as a primary publication themselves.

Since we think that CSPs represent a special subset of Complex Digital Objects, we use the term to describe an aggregation of distributed web resources relevant to scholarship.

Like any other Complex Digital Object, a CSP is considered as a logical unit with boundaries defined by the referenced resources. CSPs might be created explicitly by operators of publication systems, dynamically by automated web services or, if supplied with aggregation editing systems such as SCOPE[1], manually by researchers themselves. The component parts of an aggregation are interlinked with each other through semantic relationships. Each part is either a conceptual construct (e.g. a conference) or a more concrete entity (e.g. a text, image, video file or visualization) stored on a web server. They may have different content and semantic types, vary in their manifestations or mime types and can be distributed over different locations on the web.

So, what are the implications of these sophisticated digital objects on digital preservation, academic research, digital publishing and semantic web technologies? Before addressing these questions, it is crucial to understand how to construct, organize and exchange CSPs.

Standards for representing Complex Digital Objects

To the human eye, a website containing publications with associated materials and metadata using hyperlinks seems to be an appropriate presentation format. However, software services cannot interpret information given in a web page the same way a human reader does. For example, a search engine cannot distinguish between raw data and an instructive video that is to be found at the end of a conventional hyperlink – certainly not without additional semantic information.

In order to give human users as well as software services profound information about contextual content, Complex Digital Objects must be represented as comprehensively as possible. This requires functionality that allows for interpretation of inherent relations, file types, semantic content models and unique identification schemes (Boulal et al, 2010).

In this chapter we will briefly present two standards for data representation (OAI-ORE[2] and METS[3]) and evaluate their applicability to software services such as data exchange and Long Term Preservation.

Open Archives Initiative Object Reuse and Exchange: OAI-ORE

OAI-ORE defines a standard for the description and exchange of aggregated web resources. It introduces an abstract data model on top of the RDF model, in which each of

[1] http://www.ijdc.net/index.php/ijdc/article/download/84/55

[2] http://www.openarchives.org/ore/1.0/

[3] http://www.loc.gov/standards/mets

four main entities is represented by a Uniform Resource Identifier (URI) (Van de Sompel, 2009).

An OAI-ORE-Resource Map (ReM) describes an Aggregation (A) of Aggregated Resources (AR) and retains all information about it, including provenance information, technical and rights metadata as well as corresponding structural information. As OAI-ORE is compliant with RDF, a ReM can be serialized in different formats such as RDF/XML, Atom and N3 that makes it applicable to a broad range of data exchange on many digital platforms. Furthermore, the OAI-ORE data model supports nested Aggregations. They are able to represent Aggregations of *other* Aggregated Resources and thereby cover a typical use case in publication infrastructures for e-journals. OAI-ORE also allows the extension of its data model with third-party-vocabularies and ontologies.

Metadata Encoding and transmission Standard: METS

METS was primarily created in the digital libraries and archiving environment. It provides a framework to wrap several metadata types and, at times, an arbitrary number of file formats stored as a byte stream. A METS file can be considered a container format that typically holds descriptive, administrative, structural and technical metadata related to a digital object. A unique internal identifier references each content item in a METS section. Interlinking the sections within a METS document provides the structure of a Complex Object described in METS. The content of a section may either be stored inside the METS document or held externally and referenced from a main METS file. METS' strong abilities as a wrapper format and its capability to store binary data make it well compatible with the OAIS Reference Model[1]. It can be used for creating packages for submission (SIP), archival storage (AIP) and dissemination (DIP) within a digital archive.

The OAI-ORE and METS frameworks are both well-suited for representing Complex Digital Objects. However, they put emphasis on different purposes. While the OAI-ORE framework offers ideal tools for the interoperability of web resources within networked semantic services, METS is a convenient format for packaging binary data along with its metadata.

OAI-ORE can turn complex digital objects, such as those stored in repositories, into reusable and exchangeable web resources. As OAI-ORE relies on a standardized and flexible RDF-based data model, it became the framework of choice for tasks using the Linked Data approach. By comparison, METS appears to be an ideal preservation format. However, the preservation of dynamically changing data sets raises new questions about versioning and the idea of an 'ideal time' for a preservation snapshot.

Remember

From the definition given above, we have concluded that Compound Scholarly Publications are a subset of Complex Objects. The statement has certainly an impact on the way objects are handled in Long Term Preservation. In addition to the conventional features of web resources or other digital objects relevant for preservation, even more

[1] http://public.ccsds.org/publications/archive/650x0b1.pdf

criteria have to be considered. These include the object structure as well as the technical and semantic accessibility of each component within the Complex Object.

Compared to self-contained scholarly publications, CSPs require additional efforts in preservation. At the same time, they entail crucial improvements for the preservation of scholarly works as well as other Complex Digital Objects. (Cheung et al., 2008) note that scientific publications "*inadequately represent the earlier stages* [of the scientific process] *that involve the capture, analysis, modelling and interpretation of primary scientific data*".

This statement implies that a publication does not merely consist of final results, but rather of an aggregation of outcomes and results from the discovery process. In this sense, a publication goes through multiple stages until it becomes a "final" version. Each stage of research produces one or more outputs. Capturing these outputs in a well-structured form is likely to be interesting to a number of stakeholders (i.e. researchers, historians, archivists, authorities) concerned with the genesis of scholary works.

Compound Scholarly Publications may adapt some of the changeable nature of web resources. Used within digital research environments, they are suited for representing the discovery process close to real-time. Within the eco4r project, the actual appearance of Compound Scholarly Publications is represented within Resource Maps. They may be overwritten periodically or maintained for reuse at any desired time (Figure 1).

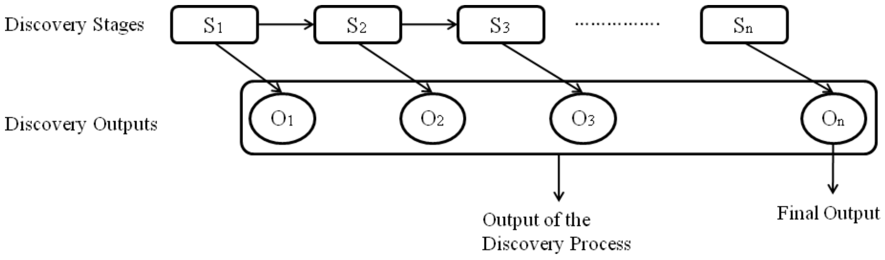


Figure 1. Discovery stages reflecting the discovery process

In order to benefit from the information provided within CSPs on the level of Long Term Preservation, some key information must be contained within them. The next section will elaborate on the minimum requirements according to good practice in digital preservation.

Relevant Features for Preserving CSPs

Descriptive Metadata

A common shortcoming of data stored in repositories is a lack of granular metadata (Boulal et al. 2010, Place et al. 2008). The "main object", for instance a text file, is usually well described, whereas supplements are hardly ever equipped with descriptive metadata. Since a CSP can include materials created by different authors stored in different repositories, the lack of metadata can bring about a proverbial 'invisibility' of data. Descriptive Metadata which are exclusively associated with a full text have no use whatsoever for supplementary materials once they reside beyond the boundaries of the repository. As an example from the eco4r project, much supplementary material is stored in repositories without explicitly describing its content. Without basic metadata such as contentual, rights and authorship information the objects are not very likely to be reused in other environments. We recommend to include the following minimum descriptive metadata for any resource used in a CSP, even if some effort must be taken to meet these recommendations in existing repositories:

- classification
- creation date
- last modification date
- authorship
- rights information.

Persistent Identifier

From the perspective of Long Term Preservation, both the integrity of digital objects and the coherence between its constituents and their persistent identification are relevant (Doorenbosch & Sierman, 2010). A global persistent identification mechanism guarantees durable validity of the relationships between the constituents as well as independence from local changes (i.e. storage location or removal) within repositories. As stated in the OAIS model, the use of stable worldwide unique identifiers is most beneficial for Long Term Preservation. Persistent identifiers allow for authentic referencing and assure a reliable assignment of metadata to the corresponding digital object. Common Internet addresses like an URL are not recommended because they usually change over time.

Moreover, the German DINI initiative[1] requires generating one persistent identifier for each object version whenever the content has changed (DINI, 2010). We recommend the use of standardized and well-established persistent identification schemes for both the aggregation (CSP) and every aggregated resource. A list of the most used persistent identification systems, specifications and standards can be found at (Neuroth et al, 2009).

Structural Information

The networked structures of CSPs can be described through semantic relationships. Place et al. (2008) classify different kinds of relationships that are able to express containments, sequences, lineages, versioning, manifestations and bibliographic citations. Many repository systems do not explicitly designate the relationships between objects (i.e. OPUS). This can lead to a major drawback. Without the ability to set relationships between objects, a CSP is not much more than loose parts of data. A piece of information such as

- “result of experiment X *is_derived_from* results of experiments Y and Z” or
- “document A *is_Annotation_of* doctoral thesis B”

is neither recorded nor processed. This way, crucial information for present and future semantic software services remains unused.

Semantic Information

In general, the semantic of a Compound Scholarly Publication is provided by a set of properties (metadata) and relationships that describe its internal and interlinking structure.

[1] http://www.dini.de/english

The CSP data model is a suitable concept from which this information can be instantiated. A data model defines classes, properties and relationships, which are applied to objects from a specific domain. Examples from the domain of scientific publishing include

- classes, i.e. "Journal_Article", Proceeding", "Thesis", "Annotation" etc.
- relationships like "has_annotation", "is_manifestation_of" etc.
- properties such as "title", "subject", "creator" etc.

A data model can be expressed in different ways, for example as an Entity Relationship Model or as an Ontology. For our purpose, we use the ontological approach since ontologies can be expressed through formal languages (e.g. OWL and RDFS) and serialized in a machine-readable form. Another useful characteristic of ontologies is the ability to deduce additional (implicit) knowledge automatically. This mechanism is known as *inference* (Allemang & Hendler, 2008).

For example, if we state that a "Thesis" is a "Publication" and that a "Master Thesis" is a "Thesis" we can deduce that a "Master Thesis" is also a "Publication" (Figure 2). For Long Term Preservation, both the explicit and implicit information in a data model can be of crucial importance. Thus, we recommend translating the underlying data representation inside the repository into ontology serializations and to archive them along with the data described.

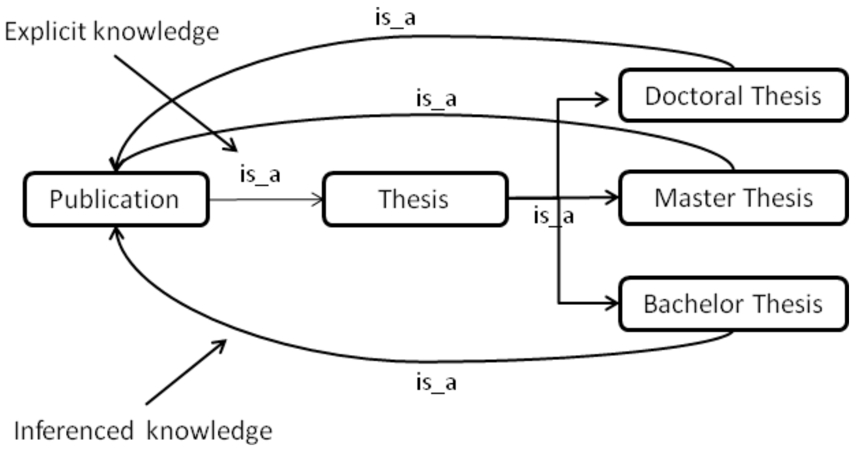


Figure 2. Example for inference

Restructure

From the scientist's perspective, reusing all kinds of information provided by a CSP is in their main interest. The ability to find, access *and* understand *all* the content of a CSP gives them the means to use the new publication form most effectively. To enable this practical reuse, we will need to find a serialization form that encapsulates key properties of CSPs to capture technical, descriptive and structural information.

Both standards discussed above are frameworks that provide tools for the construction of standardized representations of Complex Objects. However, further adaptations and refinements are needed to fairly describe objects from a specific domain.

In this section we will discuss exemplary how to use OAI-ORE in order to extract an interoperable and machine-readable representation of CSPs stored in a Fedora[1] repository.

[1] http://fedora-commons.org/

Then we will show how to include bibliographic information in an abstract OAI-ORE representation (Resource-Map).

Fedora to OAI-ORE Resource-Map

Fedora is a repository system that is well adapted for the deposit and management of Complex Objects. The underlying data model is flexible enough to depict relationships between digital objects and between different parts of an individual object as well.

Based on the Fedora data model[1], a digital object is represented as a container with different content objects known as "data streams"[2]. In addition to some reserved data streams (e.g. for storing relationships, Dublin Core[3] metadata and versioning information), they may be used to store any kind of digital material (PDFs, videos, software, visualizations, etc.). Object-to-Object relationships or relationships between content items are stored in special data streams called RELS-EXT and RELS-INT. The relational information is kept in a persistent manner in an RDF store as well.

Therefore, a mapping between Fedora and OAI-ORE data models is straightforward as shown in Figure 3.

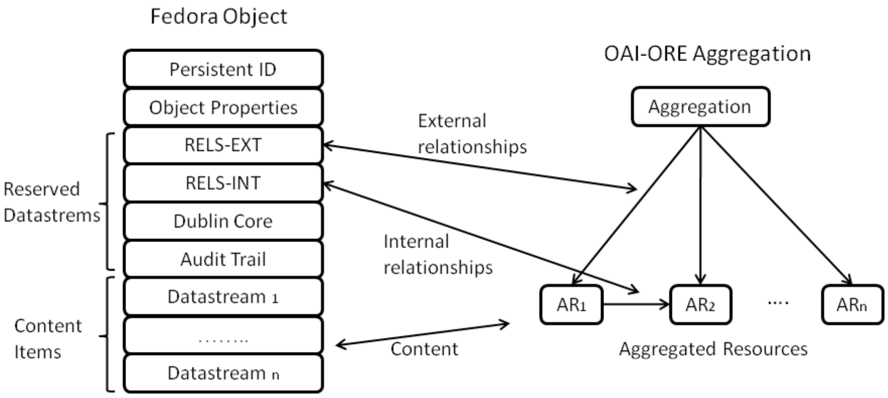


Figure 3.. Mapping between Fedora object and OAI-ORE Aggregation

Including Bibliographic Metadata in an OAI-ORE Resource-Map

As mentioned earlier, OAI-ORE defines a data model to describe aggregations of web resources. It defines an entity called Resource-Map that holds all information about an aggregation. The data model is highly generic but also expandable to enable an accurate description of objects of specific domains. OAI-ORE recommends the use of a variety of additional common vocabularies as Dublin Core or FOAF (Friend of a Friend ontology).[4]

However, to create OAI-ORE representations of Compound Scholarly Publications we will need to integrate bibliographic information in the OAI-ORE representations (Resource-Map). This can be accomplished by incorporating one of the bibliographic

[1] http://fedora-commons.org/documentation/3.0b1/userdocs/digitalobjects/objectModel.html

[2]http://fedora-commons.org/documentation/3.0b1/userdocs/
digitalobjects/objectModel.html#data

[3] http://dublincore.org/

[4] http://xmlns.com/foaf/0.1

vocabularies as The Bibliographic Ontology[1] or the FRBR-aligned Bibliographic Ontology[2] (FaBiO). The result will be an enhanced data model that is strongly based on the OAI-ORE data model and that simultaneously includes bibliographic concepts and properties for describing CSPs. Furthermore, using OAI-ORE as a basis of a data model for CSPs involves a good deal of interoperability. Every service on the web that understands OAI-ORE is able to handle CSPs as well. For example, in order to get a visualization of a CSP we can use every Resource-Map visualization service on the web (e.g. surf-incontext[3] visualization service).

The same approach has been adopted in the eco4r[4] project, where the FaBiO ontology has been used to refine the OAI-ORE data model towards a resulting data model. Figure 4 shows a simplified example of a journal article represented according to the eco4r data model. A more accurate example can be found in the project wiki[5].

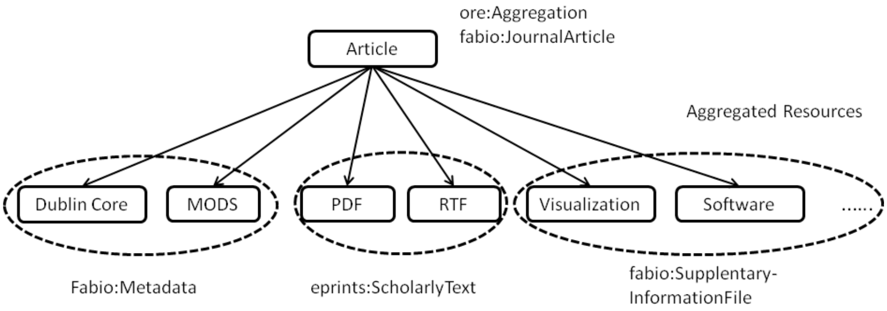


Figure4. A simplified representation of journal article according to the eco4r data model

Reuse

We define the reusability of CSPs as the ability to reuse entire publications or their component parts in other contexts. In prior sections we have seen that the process of reusing existing knowledge from a research process is crucial in order to generate new insights. Furthermore, capturing this knowledge in a well-structured form suitable for Long Term Preservation enables for durable access and reusability. We have also defined some criteria that enable CSPs to be reused by Long Term Preservation-technologies.

CSP that are represented by OAI-ORE Resource-Maps can easily be integrated in the Linked Open Data (LOD) infrastructure since they are RDF-based representations. LOD services such as lobid.org[6] (Linking Open Bibliographic Data) can reuse information about a CSP while data providers can in turn take advantage of this service to enrich their CSPs. For example, authorship information of a CSP can be enriched by connecting the corresponding Resource-Map to one or more authority services as VIAF[7] (Virtual International Authority File). Subject information like a DDC-classification that is described as a simple character (e.g. ddc:600) in a CSP can be enriched by connecting the Resource-Map to a LOD service, see Figure 5.

[1] http://bibliontology.com

[2] http://speroni.web.cs.unibo.it/cgi-bin/lode/req.py?req=http:/purl.org/spar/fabio

[3] http://code.google.com/p/surf-incontext

[4] http://www.eco4r.de

[5] http://trac.eco4r.org/trac/eco4r/wiki/DataModel

[6] http://lobid.org/de

[7] http://viaf.org

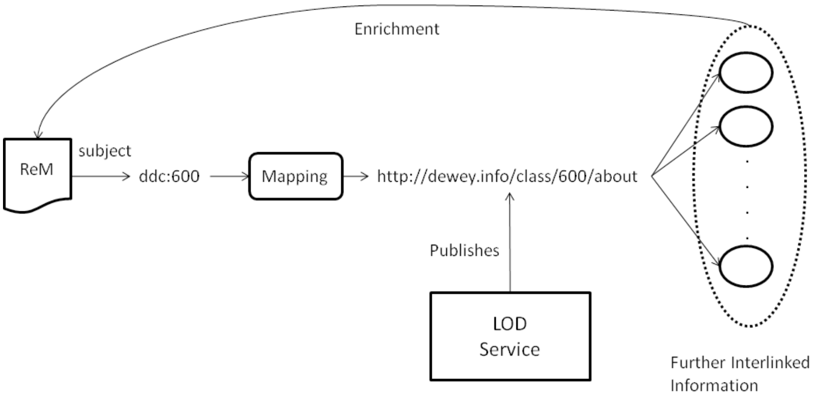

Enrichment
ReM
subject
ddc:600
Mapping
http://dewey.info/class/600/about
Publishes
LOD Service
Further Interlinked Information

Figure 5. Metadata enhancement with Linked Open Data Services

Although OAI-ORE requires common vocabularies such as Dublin Core together with domain specific information, it does provide a data model that may be used as a generic interoperability layer by different repositories. For example, in the eco4r project, we show that Resource-Maps can be collected dynamically from different repositories in order to build new aggregations according to the DDC[1]-classification. By using a common data model (OAI-ORE), the integration and processing of CSP representations by a presentation layer[2] are both straightforward.

Case Study

With respect to its practical approach, the project's first effort was to analyse real life publication material stored in two different repository systems. Both Fedora and OPUS are hosted at the projects partners, hbz NRW[3] and Bielefeld University Library[4], respectively. We took the results from this evaluation to define practical requirements for an implemented prototype system that exposes and reuses real life CSPs. The system follows the principle of an "Overlay Journal". Technically, it will demonstrate the ability to gather information resources (including CSPs) from existing repository systems as well as the dynamic aggregation and creation of new CSPs on top of the gathered resources. In consequence, the Overlay Journal gives access to aggregated and newly arranged pieces of information for human researchers as well as automated systems. If metadata and semantic information are provided sufficiently, the information segments remain useful regardless to their actual composition.

In more detail, the prototype system collects OAI-ORE Resource Maps representing CSPs stored in the source repositories. After storing the collected representations in a RDF store, the Overlay Journal performs a sequence of processing steps. They include metadata enhancement through Linked Open Data services and the creation and exposure of new thematic aggregations. Figure 6 gives an overview about the architecture of the Overlay Journal.

[1] http://www.oclc.org/dewey

[2] In the eco4r project the presentation layer is an Overlay Journal

[3] http://www.hbz-nrw.de

[4] http://www.ub.uni-bielefeld.de

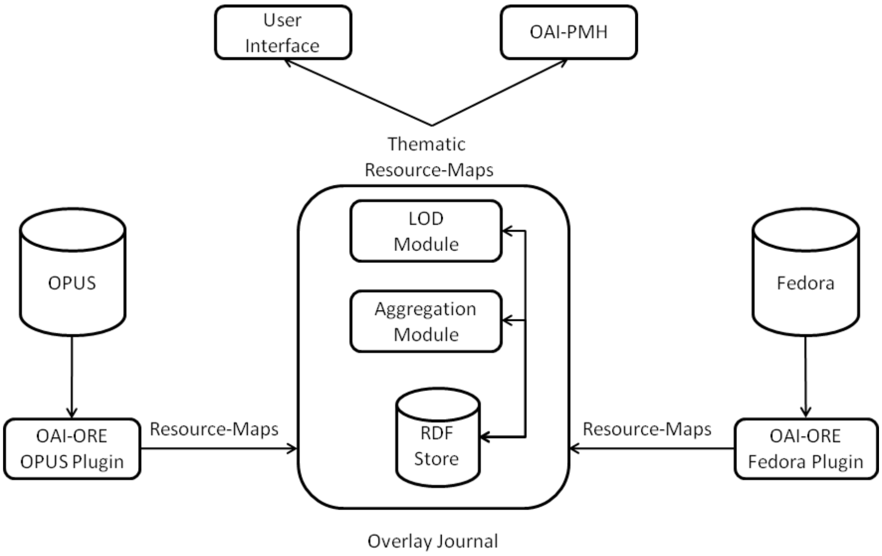


Figure 6:. Overlay Journals architecture and repository interaction

The Repositories

The repositories used in eco4r are quite different in terms of the software systems used and the materials stored within them. The Bielefeld University Library uses the OPUS[1] repository software. Doctoral theses are the predominant content followed by some post-prints of published articles. The Library Service Centre in Cologne (hbz) operates and manages a publishing infrastructure for Open Access journals[2], wherein a Fedora[3] repository system stores the digital objects, representing journal articles and various supplementary materials (see Boulal et al. 2010).

The OAI-ORE Repository Plugins

The main practical outcomes from the project are OAI-ORE software plug-ins for each of the repository systems. Their function is to generate OAI-ORE Resource-Maps from the repositories' content according to the recommendations given before. Nevertheless, both plug-ins are designed to be generic and independent from the specific applications and data models used on top of Fedora and OPUS. By using a configuration mechanism, adjustments can be made to determine exactly how the generated Resource-Maps should look like. For example, the Fedora plug-in can be configured to select metadata, relationships and resources that should appear in the Resource-Map. Furthermore, a mapping mechanism is implemented to substitute repository-specific data structures with globally standardized relationships such as DC terms relationships, the FRBR-based FABIO ontology or the FOAF namespace. This way, we have harmonized the need for a deployment of standardized metadata with the practical requirements of proprietary repository systems.

[1] http://www.opus-repository.org/index.html

[2] http://www.dipp.nrw.de

[3] http://fedora-commons.org

The OAI-PMH Harvester

In practice, the OAI-ORE Resource-Maps are exposed through an OAI-PMH Interface. This way, an OAI Harvester can easily communicate with the repository's OAI-ORE plug-in. In our case, the harvester exposes the generated Resource-Maps as RDF/XML serialized items to the business logic of the Overlay Journal. By using the OAI-Protocol we introduced a rather simple but very robust technique to manage – i.e. find, harvest and update – CSPs in repository systems.

The RDF Store

Any Resource-Map provided by the Overlay Journal is stored in a RDF Store. In eco4r we use for the software Sesame 2.4.

The Linked Open Data (LOD) Module

With the LOD module, we aim to show that the information stored in CSP can be enhanced when connecting the associated Resource-Map with Linked Open Data services. In eco4r we are connecting to the lobid.org[1] service. The LOD module fetches the RDF store periodically and attempts to find intersections with information published in lobid.org. For example, in a Resource-Map stored in the RDF store we can find subject information according to the DDC[2] classification scheme, such as ddc:600 which represents the category "Technology". The LOD module fetches lobid.org and finds a resource for this category (e.g http://dewey.info/class/600), which is in turn connected to other information resources. The module will update the information stored in the RDF store by replacing the character "ddc:600" with the URI "http://dewey.info/class/600/". Thus, a metadata enhancement has been performed (Figure 5).

Aggregation Module

The Aggregation module is the second use case to demonstrate the reusability of Resource-Maps. Here, new Aggregations are generated according to thematic properties such as the DDC classification. These new Aggregations will then be visualized in a user interface and, in turn, exposed via OAI-PMH.

Conclusion and Outlook

Scholarly communication is undergoing a major change. The emergence of modern technologies is opening new possibilities to the way scientists perform and disseminate their research. However, current publication infrastructures still focus on processing single monolithic resources within isolated data silos. They are not able to satisfy the

[1] http://lobid.org/de

[2] http://www.oclc.org/dewey/about/default.htm

growing demand for interlinked information resources that connect existing research data infrastructures using Semantic Web technologies.

For example, the High Level Expert Group on Scientific Data[1], which was assigned to prepare a “Vision 2030” [2] for the evolution of e-infrastructure for scientific data, recommends a larger degree of integration within scientific infrastructures. It states that the emerging Linked Data technologies have the potential to fulfil this requirement (Bizer, 2011).

Compound Scholarly Publications are a good example to illustrate a practical use case for linking different data resources using Linked Data and Semantic Web technologies. In the eco4r project as well as in associated projects, many theoretical and practical results have been achieved. However, we have discovered a lack of practical and real-life applications that allow for the processing of Complex Objects in scientific environments.

Having presented an executable technical approach for restructuring and reusing these valuable resources, we encourage both repository managers and data curators to affiliate their data pools as much as possible. Using tools and frameworks for data integration and Open Access publishing will not only revalue today’s data but also ensure its continued existence.

Acknowledgments. The eco4r project is a German Research Foundation (DFG) funded collaboration between Bielefeld University Library and North Rhine-Westphalian Library Service Centre (hbz). We would like to thank our Bielefeld colleagues Wolfram Horstmann, Jochen Schirrwagen and Friedrich Summann for their specialist support before and during the project. We also owe our thanks to the international participants of the 2010 eco4r workshop, which brought about very valuable insights into the way complex digital objects are – and should be – handled in future technical environments.

References

Allemang, D., Hendler, J. (2010). Semantic Web for the Working Ontologist – Modeling in RDF, RDFS and OWL

Bizer, C. (2011). *Expert Report on Linking Data & Publications*. Freie Universität Berlin.

Boulal, A., Quast, A., Iordanidis, M., & Schirrwagen, J. (2010). *Report on Enhancing Interoperability between existing Open Access Publication Infrastructures*. Library Service Center NRW, Cologne & Bielefeld University Library, Bielefeld.

Cheung, K., Hunter, J., Lashtabeg, A., & Drennan, J. (2008). *SCOPE – A Scientific Compound Object Publishing and Editing System*. The University of Queensland, St Lucia, Queensland, Austria.

DINI, Deutsche Initiative für Netzwerkinformation e.V., Arbeitsgruppe Elektronisches Publizieren (2010): DINI Certificate – Document and Publication Services 2010. http://nbn-resolving.de/urn:nbn:de:kobv:11-100182800

Doorenbosch, P., & Sierman, B. (2010). *Institutional Repositories, Long Term Preservation and*

the changing nature of Scholarly Publications. Koninklijke Bibliotheek, the Netherlands.

Maslov, A., Creel,J. Mikeal,A. Phillips, S., Leggett, J. (2009).Adding OAI-ORE Support to Repository Platforms URI http://hdl.handle.net/1853/28448

Neuroth, H; Oßwald, A.; Scheffel, R.; Strathmann, M.; Huth, K. (2009). nestor Handbuch – Eine Enzyklopädie der Digitalen Langzeitarchivierung

Place, T., Van der Feesten, M., Hoogerwerf, M., Bijsterbosch, M., Slabbertje, M., Hogenaar, A., et al. (2008). *Report on Object Models and Functionalities – DRIVER II*. Leiden University.

Ruijgrok, P., Slabbertje, M., & Van Luijt, M. (2009). *A candidate semantic representation for enhanced e-theses: Guidelines for modeing enhanced e-theses*. Knowledge Exchange – University of Utrecht.

Van de Sompel, H. (2009). The OAI-ORE Interoperability Framework in the Current Scholarly Communication Context. Los Alamos National Laboratory, USA.

[1] http://cordis.europa.eu/fp7/ict/e-infrastructure/high-level-group_en.html

[2] Charged by the European Commission’s Directorate-General for Information Society and Media.

Biographical sketches for lead authors (in order of appearance)

Professor Richard Beacham, King's College London, is a native of Virginia. He earned his BA and doctorate at Yale University. At Yale College he studied ancient history and classics, and earned his DFA at the Yale School of Drama in dramaturgy, theatre history, criticism, and dramatic literature. He has made his home in the UK since 1974. From 1976 until 2005 he worked at the University of Warwick, where he co-founded its School of Theatre Studies. He has also worked as a visiting Professor at Yale and at the University of Santa Barbara, and spent a year as a "Museum Scholar" at the Getty Center in Los Angeles. He has translated and published comedies by Plautus, and produced and directed these. He and the research team he leads, the King's Visualisation Lab, came to King's in 2005. He is a leading international authority on the use of the computer based research and 3D visualisation of historical buildings and artefacts.

Drew Baker is a Research Fellow within the Department of Digital Humanities, Kings College London. One of the founding members of the King's Visualisation Lab he has worked in the field of 3D visualisation and interpretation of archaeology and history since 1997. He has specialised in the area of 3D modelling specifically using interactive VRML and virtual world technologies. His primary area of interest is in using 3D and advanced technology to bring cultural history from traditional passive media into interactive new media transforming the user into an active participant though exploration of virtual worlds and artefacts, the process of developing such environments and interactions and the long term preservation of digital cultural heritage. His personal research area is in the 1st century Campania region of Italy focusing on the Roman colony of Pompeii, insula VIII.7 and the history of the rediscovery and archaeological excavations of the area since their destruction in 79AD.

Dr. David Anderson is CiTech Research Centre Director, and co-leader of the Future Proof Computing Group at the University of Portsmouth. He holds a B.A. Hons in Philosophy (QUB) and a Ph.D. in Artificial Intelligence (QUB). His research interests include: digital preservation, history of computing, paraconsistent logic, epistemology, artificial intelligence and the philosophy of mind. He is the UoP Principal Investigator for the EC FP7 Project KEEP. David has served on numerous international committees including the IEEE Publications committee. He is a member of the AHRC Peer Review College.

Dr. Janet Delve is co-leader of the Future Proof Computing Group, one of the research clusters in the Centre for Cultural and Industrial Technologies Research (CiTech) at the University of Portsmouth. She holds degrees in Mathematics (UCL), French (Southampton), together with a Masters degree in Microwaves and Modern optics (UCL), and a PhD in the History of Mathematics (Middlesex University). Her research interests include: metadata modelling for digital preservation; data warehousing applied to cultural domains; and the crossover between the history of computing and digital preservation. The University of Portsmouth is a partner in the EC FP7 Project KEEP, in which Janet is responsible for the data modelling of complex digital objects and the development of the technical environment metadata database, TOTEM. She is a member of the AHRC Peer Review College.

Neil Chue Hong completed an MPhys degree in Computational Physics from the University of Edinburgh. He is the PI and Director of the EPSRC-funded Software Sustainability Institute, a national facility based at the University of Edinburgh for research software users and developers providing specialist software engineering skills to drive the continued improvement and impact of research software. His current research interests are in community engagement and development, software sustainability, and the integration and analysis of data. He works with research communities across the UK and globally to promote and enable the improvement of important research software through consultative advice; collaborative partnerships; and long-term engagement. From 2007-2010, he was Director of OMII-UK, based at the University of Southampton, which provides and supports free, open-source software for the UK e-Research community. Neil has also worked extensively with Scottish SMEs, primarily on database and image processing projects, in areas as diverse as: fishing net simulators, drilling control software, heart analysis, image archiving, semiconductor probe inspection, genetic expression in mice, and mushroom sorting.

Dr. Brian Matthews is Group Leader of the Scientific Applications Group within the e-Science Centre, and also Deputy Manager of the W3C Office for the UK and Ireland. He has worked on a wide-range of different areas of computing and information technology. He took his PhD on equational reasoning systems with Glasgow University, and worked in the area of formal methods for software engineering. Brian then worked on projects on structured documentation, data modelling and the Web. He has been involved with the Semantic Web, developing early versions of the RDF format for thesuari and other knowledge organisation systems, which has evolved into the *SKOS recommendation*. In recent years, Brian has become increasingly involved in developing technology and tools to support scientific research and collaboration. This has involved work in data management, preservation and distributed systems. He has also contributed to developing STFC's institutional repository ePubs, and has been involved in other projects in digital libraries, information management, and digital curation. Brian has also worked with distributed systems and Grids, including developing tools and techniques for managing security and trust in Grids. He has been the STFC leader in the European Project XtreemOS, which is developing a Grid based operating system.

Dr. Arif Shaon has provided technical leadership to numerous projects in the area of advanced environmental informatics and digital preservation at the e-Science centre of the Science and Technology Facilities Council (STFC). Dr. Shaon's past notable endeavours include the development of preservation approaches for geospatial data and software artefacts, and the integration of the UK National Grid Service with OGC data processing services (as listed in the JISC Standards catalogue). He also has considerable experience in the emerging areas of Semantic Web, particularly in terms of developing linked-data approaches to publishing geospatial data in order to support the UK Open Data initiative. At present, he is playing leading roles in a number of preservation focused EU projects. He is also a member of the EuroSDR working group — an initiative comprising digital preservation experts and archivists from various national mapping agencies and archives in Europe collaborating to address the issues of long-term digital preservation of European geospatial data. In addition, he holds a Ph.D. in long-term metadata curation as applied to long-term digital preservation.

Esther Conway is an Earth Observation Data Scientist at the Centre for Environmental Data Archival. She is an experienced researcher and analyst in the area of digital preservation, having worked on a variety of EU and UK based research projects. Her main area of research has been the development of methods and models to support the long term exploitation of scientific research assets. Esther originally trained in Physics at Imperial College London and subsequently acquired a Masters in Information Systems and Technology from City University London.

Professor John Clarke, University of Texas, received his Ph.D from Yale University. In 1980 he began teaching at The University of Texas at Austin, where his teaching, research, and publications focus on ancient Roman art, art-historical methodology, and contemporary art. His work has focused on the visual culture of ancient Rome, on art-historical methodology, and on contemporary art and criticism. He has published seven books; two appeared in 2007: Looking at Laughter: Humor, Power, and Transgression in Roman Visual Culture, 100 B.C.-A.D. 250 (University of California Press) and Roman Life: 100 B.C.-A.D. 200 (Abrams). He has published numerous articles, chapters, and reviews, including several on the mosaics and paintings of Villa A at Oplontis, Torre Annunziata, Italy. Currently Clarke is co-director of the Oplontis Project, a collaboration with the Archaeological Superintendency of Pompeii and the King's Visualisation Lab, King's College, London. The Oplontis Project will furnish a comprehensive publication of this huge luxury villa (50 B.C.-A.D. 79), with all the research findings keyed to a navigable, 3D digital model. Support for the project includes a Collaborative Research Grant from the National Endowment for the Humanities.

Dr. Kenton McHenry, NCSA University of Illinois at Urbana-Champaign, Dr. Kenton McHenry received a Ph.D degree in computer science from the University of Illinois at Urbana-Champaign in 2008 after completing a B.S. degree from California State University of San Bernardino. He is currently a Research Scientist at the National Center for Super-computing Applications (NCSA) and lead of the Image, Spatial, and Data Analysis (ISDA) group. His background is in computer vision with interests in the areas of image segmentation, object/material recognition, and 3D reconstruction. At NCSA he has applied this experience towards the task of digital curation, specifically, digital preservation and access. In recent years he has worked on a series of tools focused around the need for file format conversions in digital archives. One such tool, NCSA Polyglot, is a file format conversion service that was designed as an extensible, distributed, and practical solution to accessing data among the many contemporary (and legacy) file formats available. The Polyglot service is built on top of another tool developed by the ISDA group, the "software server". Software servers are background processes run on a machine to provide API-like access to installed applications. These servers essentially turn traditional desktop software into web based services and are what allow Polyglot to carry out a potentially large number of format conversions.

Jenny Mitcham, The Archaeological Data Service, The University of York, was trained as a field archaeologist but after a few years working on excavations at home and abroad, decided that it would be kinder to her wrists and knees to move into the sphere of IT. After a MSc course in archaeological computing, she worked in several different places including as a Sites and Monuments Officer before eventually moving to the Archaeology Data Service (ADS) where she has been since 2003. One of the core functions of the ADS is to archive digital materials for the long term and make these objects available for reuse. She has been learning digital archiving on the job for the last 8 years, facing the wide range of challenges that archaeological data provides and has a particular interest in standards and certification in digital archiving and re-usability of data. The Archaeology Data Service (ADS) works to preserve and disseminate the wide range of data types that archaeologists produce. These range from simple text reports, spreadsheets and digital photographs to the far more complex 3-dimensional digital models. Although, for the ADS, the principles of digital archiving remain the same no matter what the data, complex data, which is often large in size and in proprietary formats brings new challenges.

Dr. Hugh Denard, King's College, London earned his BA in Theatre and Classical Civilizations at Trinity College, Dublin in 1992. He completed an MA in Ancient Drama and Society in the Classics Department at the University of Exeter in 1993 and a Ph.D on versions of Greek Tragedy by Irish writers, in the Department of Drama at Exeter, in 1997. He held a one-year Teaching Fellowship in the English Department at Trinity College Dublin (1997-8), before moving to the School of Theatre Studies at the University of Warwick in 1998, where he taught students of Theatre and Performance Studies, coordinated the Theatre, Media and Text degree and worked as a researcher on the AHRC Theatre of Pompey Project. Hugh moved to King's College London, with the other members of the Visualisation Lab, in September 2005. From January to March 2011, he was Visiting Research Fellow at the Long Room Hub, Trinity College Dublin, where he studied the early Abbey Theatre and produced a research-based, mixed media performance at the Samuel Beckett Theatre.

Daniël Pletinckx was trained as a civil engineer, with specialization in information technology. He gained extensive experience in system design, quality assurance, digital image processing and synthesis, 3D and virtual reality through a career of 15 years in private industry. He is the author of several articles on computer graphics and cultural heritage presentation and has lectured extensively at major computer graphics and cultural heritage conferences. Daniël Pletinckx was chief consultant to the Ename 974 project, a major heritage project in the historical village of Ename, Belgium, and founded the international Ename Center for Public Archaeology and Heritage Presentation, together with Dirk Callebaut and Neil Silberman. He designed four TimeScopes in Ename, which received three international awards (Golden Scarab and Flemish Monument Award in 1998 for TimeScope1, the VGI ICT Award for best innovation in public outreach in cultural heritage for TimeScope3). Currently, Daniël Pletinckx is director of Visual Dimension bvba, a SME dealing with consulting on and designing of new systems for cultural heritage and tourism. Visual Dimension consults major European heritage organisations on innovating and optimising the use of ICT technology in tangible and intangible cultural heritage and tourism. Visual Dimension also specialises in new, efficient ways to digitise existing cultural heritage objects and monuments, and in virtual reconstruction of historical buildings and landscapes.

Andrew Ball, Former head of IT Audit, The Audit Commission, is an experienced IT audit professional who spent twelve years with the UK Audit Commission. His professional background is in IT service, project and programme management with over 20 years experience. He holds a number of professional qualifications in PRINCE2, ITIL and MSP. Andrew is also a Certified Information Systems Auditor and a Chartered Member of the British Computer Society.

Clive Billenness holds the post of EC Project Management Expert at the British Library. He is currently the Project Manager for POCOS and also a workpackage lead on the EC FP7 Project KEEP as well as a member of the British Library's project team on the EC FP7 Project SCAPE. Qualified in Prince2, MSP and M_o_R, Clive was the Programme Manager of the Planets Project and a member of the team which created the Open Planets Foundation. He is a Certified Information Systems Auditor. As a Head of the Northern Region's Public Sector Information Risk Management Team at KPMG LLP, Clive was responsible for directing a review on behalf of the National Audit Office of the £30m project to update the Department of Work and Pensions computer systems. He also advised the UK Office of Government Commerce and the Office of the UK Deputy Prime Minister on a number of IT Projects. Prior to this, Clive was a Regional Service Lead for the Audit Commission where he was frequently loaned to clients to assist with the recovery of projects which were in exception. Clive is a member of the Office of Government Commerce's Examining Board for Project, Programme and Risk Management examinations. He is also a Director of the UK Best Practice User Group for the same disciplines. Clive is a regularly published author for the Chartered Institute of Public Finance and Accountancy (CIPFA) on Project Management.

Anouar Boulal studied computer science and economics with the focus on data mining and information retrieval at the University of Bonn, Germany. After a constitutive work experience at the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), he started working as scientific assistant and Java developer at hbz Library Service Center, Cologne. At hbz, he was engaged in the project '*eco4r* - Exposing Compound Objects for Repositories' and was responsible for the design and implementation of the back-end project software. Anouar Boulal is currently working as software developer in digital marketing and the information analysis sector at Experian Deutschland GmbH.

Martin Iordanidis received his M.A. degree in humanities computing, musicology and English literature at University of Cologne, Germany in 2003 focusing on the data modelling capabilities of XML Schema. He was engaged in several research projects accompanying digitisation efforts in Germany and started working as a repository manager at hbz Library Service Centre, Cologne in 2004. Since 2008 Martin Iordanidis is working as Preservation Manager at hbz and focuses on international research efforts the field of digital preservation. He is active in various European research communities such as the German nestor network and the Digital Preservation Coalition. Martin is currently enrolled in a Masters Program in Library and Information Science (LIS) and works as a rock journalist and live musician after darkness.

Andres Quast studied geology in his 'pre-librarian life'. Andres received a doctoral degree in geosciences. His new life started in 2005 with a geosciences-related project (GEO-LEO, Virtual Library for Earth Sciences and Astronomy) at the Lower Saxony State and University Library in Göttingen (SUB). He has continuously worked in different fields and projects touching the areas of Open Access, Repository Systems and Digital Preservation in Göttingen and, later on, at the North Rhine Westfalian Library Service Center (hbz) in Cologne. Since 2010 Andres is leading the ePublishing Systems department at hbz.

Glossary

Access: the process of turning an *AIP* into *DIP,* ie using data from a digital archive

ADF Opus: A Microsoft Windows–based program to create *ADF*

ADF: Amiga Disk File, a file format used by Amiga computers and emulators to store images of disks

ADS: Archaeology Data Service, a digital archive specialising in archaeological data based in York

AHDS: Arts and Humanities Data Service, a data service for higher education, closed in 2008

AIMS: Project funded by Mellon foundation to examine archival principles in the digital age

AIP: Archival Information Package, a package of information held within an *OAIS*

APA: Alliance for Permanent Access, a European network, set up *APARSEN*

APARSEN: a Network of Excellence funded by the *EC*, see *APA*

API: an interface provided by a software program in order to interact with other software applications

Archival Storage: The *OAIS* entity that contains the services and functions used for the storage and retrieval of *AIP*

ARCOMEM: ARchive COmmunities MEMories, *EC*-funded project in digital preservation

ASCII: American Standard Code for Information Interchange, standard for electronic text

BADC: British Atmospheric Data Centre

BL: British Library

BlogForever: *EC*-funded project working on robust digital preservation, management and dissemination facilities for weblogs

BLPAC: British Library Preservation Advisory Centre – a service of the BL which promotes preservation

BS10008: a British standard pertaining to the evidential weight of digital objects

CCSDS: Consultative Committee for Space Data Systems, originators of the *OAIS* standard

CD-ROM: Compact Disc, read-only-memory

Characterisation: stage of ingest processes where digital objects are analysed to assess their composition and validity

Checksum: a unique numerical signature derived from a file. Used to compare copies

CiTech: Centre for Cultural and Industrial Technologies Research

Cloud (cloud-computing, cloud-based etc.): on demand, offsite data storage and processing provided by a third party

CRT: Cathode ray tube

CSP: Compound Scholarly Publication

CVS: Concurrent Versions System or Concurrent Versioning System, a client-server revision control system used in software

Data Dictionary: A formal repository of terms used to describe data

DCC: Digital Curation Centre, data management advisory service for research

DDC: Dewey Decimal Classification

Designated Community: group of users who should be able to understand a particular set of information

DigiCurVE – Digital Curation in Vocational Education, assessment project funded by EU on training provision in Europe

Digital Object: a set of bit sequences, e.g. a single document such as a PDF file, or an image of a (console) game, etc.

DIP: Dissemination Information Package, the data disseminated from an *OAIS*

DOS: Disk Operatins System

DP: Digital preservation

DPA: Digital Preservation Award, biannual prize awarded by the *DPC*

DPC: Digital Preservation Coalition, a membership body that supports digital preservation

DPTP: Digital Preservation Training Programme, an intensive training course run by *ULCC*

DRIVER: Digital Repository Infrastructure Vision for European Research

DROID: tool developed and distributed by TNA to identify file formats. Based on *PRONOM*

DSA: Data Seal of Approval, a process by which organisations can undertake self-evaluation of their DP practices

DVD: Digital Versatile Disk, formerly the same abbreviations was used for Digital Video Disk

EC: European Commission

Edina: a national data centre based in Edinburgh University mainly funded by JISC

Emulation Framework: a framework that offers emulation services for digital preservation

Emulation: adapts a computer environment so that it can render a software artefact as if it were running on its original environment

Encapsulation: a process where digital objects are captured with information necessary to interpret them

ENSURE: Enabling kNowledge Sustainability Usability and Recovery for Economic value, *EC*-funded project

EPSRC: Engineering and Physical Sciences Research Council, UK

EU: The European Union

FOAF: Friend of a friend, machine-readable ontology describing persons

FRBR: Functional Requirements for Bibliographic Records

GD-ROM: Giga Disc Read Only Memory, proprietary optical storage medium for the game console Sega Dreamcast

GIF: Graphic Interchange Format, an image which typically uses *lossy compression*

GIS: Geographical Information System, a system that processes mapping and data together

HATII: Humanities Advanced Technology and Information Institute at Glasgow University

HDD: hard disk drive

HEI: Higher Education Institution

HTML: Hypertext Markup Language, a format used to present text on the World Wide Web

IGDA: International Game Developers Association

Incremental: a project funded by *JISC* at HATII and Cambridge University

Ingest: the process of turning an *SIP* into an *AIP,* ie putting data into a digital archive

ISO: International Organization for Standardization, body that promotes standards

JISC: Joint Information Systems Committee of the Higher Education Funding Councils

JPEG 2000: a revision of the *JPEG* format which can use *lossless compression*
JPEG: Joint Photographic Experts Group, a format for digital photographs which is *lossy*
KB: Koninklijke Bibliotheek, national library of the Netherlands, partner in *KEEP* and *APARSEN*; *APA* home to *LIBER and NCDD*
KEEP: Keeping Emulation Environments Portable, EC-funded project to develop *emulation* services to run on a virtual machine
KVL: King's Visualisation Lab
LC: Library of Congress
LCD: Liquid Crystal Display
LED: light emitting diode
LIBER: network of European Research Libraries involved in *APARSEN* and *AP,* offices at the *KB*
LIDAR: Light Detection And Ranging, an optical remote sensing technology used to measure properties of a target using light or laser.
LiWa: Living web archives, *EC*-funded project which developed web archiving tools
LOCKSS: Lots of Copies Keeps Stuff Safe a DP principle made into a toolkit for E-Journal preservation, see *UKLA*
LOD: Linked Open Data
Lossless compression: a mechanism for reducing file sizes that retains all original data
Lossy compression: a mechanism for reducing file sizes which typically discards data
MANS: Media Art Notation System MANS
Memento: an innovative tool which allows time based discovery of web pages, winner of *DPA* 2010
METS: Metadata Encoding and Transmission Standard, a standard for presenting metadata
Migration: the process of moving data from one format to another
MLA: Council of Museum Libraries and Archives, strategic body for such organisations in England
MP3: digital audio format (standing for both MPEG-1 or MPEG-2 Audio Layer III)
NARA: US National Archives and Records Administration
NCDD: Dutch national digital preservation coalition, closely aligned with *APA, DPC* and *Nestor* and hosted by *KB*
NDAD: UK National Digital Archive of Datasets, formerly funded by *TNA* and operated by *ULCC*
NDIIPP: National Digital Information Infrastructure and Preservation Programme – a major programme from the *LC*
Nestor: German network of expertise in digital preservation, closely aligned to APA and NCDD
NRW: North Rhine-Westphalia, state of Germany
OAI-ORE: Open Archives Initiative Object Reuse and Exchange, standards for description and exchange of web resources.
OAI-PMH: Open Archives Initiative Protocol for Metadata Harvesting
OAIS: Open Archival Information System, a reference model describing a digital archive
OCLC: Online Computer Library Center, Inc., US-based library and research group
OMII-UK: open-source organisation that empowers the UK research community by providing software for use in all disciplines of research
Open source: software in which the underlying code is available for free
OPF: Open Planets Foundation**,** a membership organisation which sustains outputs from the *PLANETS* project

OSS: Open Source Software

Paradata: Information about human processes of understanding and interpretation of data objects, e.g. descriptions stored within a structured dataset of how evidence was used to interpret an artefact.

PARSE.INSIGHT: EC-funded project that developed a roadmap for DP infrastructure in Europe

PDF/A: a version of the PDF standard intended for archives

PDF: Portable Document Format, a format for producing and sharing documents

PLANETS: a project funded by the EC to develop a suite of DP tools including *PLATO*. Now maintained by *OPF*

PLATO: a *preservation planning* tool which was created by the *PLANETS* project

PNM: Preservation Network Model

POCOS: Preservation Of Complex Objects Symposia, a JISC-funded project which organised a series of three symposia on preservation of Visualisations and Simulations; Software Art; and Gaming Environments and Virtual Worlds in 2011-12

PREMIS: Preservation Metadata: Information Strategies, metadata standard

Preservation planning: defining a series of preservation actions to address an identified risk for a given set of *digital objects*

PrestoPRIME: *EC*-funded project which develops tools and services for the preservation of digital audio-visual content

PRONOM: a database of file formats with notes on associated issues. Used with *DROID*

PROTAGE: Preservation organizations using tools in agent environments, *EC*-funded project

PSD: Adobe PhotoShop file format

RCUK: Research Councils UK

RDF: Resouce Decription Framework

RIN: Research Information Network, a group that studies and reports on research needs

RLG: Research Libraries Group, US research group that produced *TDR*. Now part of *OCLC*

RLUK: Research Libraries UK

SaaS: software as a service, architecture whereby software is managed remotely by a service provider (see also *cloud*)

SCAPE: Scalable Preservation Environments, *EC*-funded project developing scalable preservation actions

SHAMAN: Sustaining Heritage Access through Multivalent Archiving, *EC*-funded project

Significant properties: concept whereby identifying the most important elements element of a file will aid preservation

SIP: Submission Information Package, data received into an *OAIS*

SKOS: Simple Knowledge Organization System, specivications on knowledge organisation system, developed by W3C

SPEQS: Significant Properties Editing and Querying for Software

SSMM: Software Sustainability Maturity Model

STFC: Science and Technology Facilities Council, UK

STM: Science Technology and Medicine – major area of publishing**,** sometimes meaning the STM Publishers Association

SWISH: joint venture between *RCAHMS* and *RCAHMW* to provide digital services including long term preservation

TDR: Trusted Digital Repository, a standard which characterises ‘trust’ in a digital archive

TIFF: Tagged Image File Format, a common format for images typically *lossless*

TIMBUS: an EC-funded project which is investigating the preservation of online services

TOTEM: Trustworthy Online Technical Environment Metadata Database

TRAC: Trusted Repository Audit and Certification, toolkit for auditing a digital repository

UBS: universal serial bus

UKDA: UK Data Archive University of Essex, digital archive for social and economic data

UKLA: UK *LOCKSS* Alliance, a service of *Edina* which offers E-journal preservation

UKWAC: UK Web Archiving Consortium

ULCC: University of London Computer Centre, host of *NDAD* and creators of *DPTP*

UMD: Universal Media Disc; proprietary CD-ROM format of Sony Computer Entertainment

UML: an industry standard for visualisation, specification construction and documentation of artefacts of software systems

UNESCO: United Nations Educational, Scientific, and Cultural Organization: an agency of the United Nations supporting programmes to promote education, media and communication, the arts, etc.

VHS: Video Home System, videocassette recording technology

Virtualization: creation of a virtual rather than actual instance of software or hardware (see also *emulation*)

VRML: Virtual Reality Modelling Language, file format for representing 3D graphics

W3C: World Wide Web Consortium

WF4EVER: Advanced Workflow Preservation Technologies for Enhanced Science, *EC*-funded project

WinUAE: Amiga emulator supporting 5.25” and 3.5" double density disks and 3,5” high density floppy disks

XML: Extensible Markup Language, a widely used format for encoding information

POCOS

Preservation of Complex Objects Symposia

Project partners:

KING'S College LONDON

BRITISH LIBRARY

JOGUIN

University of Glasgow

University of Portsmouth

ISBN 978-1-86137-6305