# Planning for the End from the Start: An Argument for Digital Stewardship, Long-Term Thinking and Alternative Capture Approaches for Digital Content

**Somaya Langley**

**Abstract** Sustainability and continued access to digital cultural heritage, digital humanities content and research materials can be challenging. For any research project, available resources and dependencies set the limits for what is possible. In the digital environment, consideration of these limitations can tend to focus on the technological aspect. However, it is not just technology that ensures the success of a project or long-term access to digital content. Using the Three-Legged Stool Model for Digital Preservation (Kenney and McGovern in Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems, 2003 [38]) (and other relevant models) provides an important foundation to ensure that any digital cultural heritage or digital humanities project is approached holistically. In addition, digital stewardship (Lazorschak in The Signal, 2011 [44]) should also be considered as an essential building block for digital cultural heritage and the digital humanities. Historically, questions of sustainability and ongoing access are often brought to the fore only as funding streams near their end, or as research project champions retire. Sustainability of digital content has been a topic of debate for many years (Bodleian Libraries in Digital Humanities Archives for Research Materials, Oxford, [2], Cantara in Longterm Preservation of digital humanities in OCLC Systems and Services 22:38–42, 2006 [10]). In recent years, the importance of sustainability is being further recognised, with research funding bodies requiring plans for long-term preservation and access as a part of applications for project funding, such as requiring the inclusion of this information in Data Management Plans (DMP) (UK Research and Innovation—Arts and Humanities Research Council in Research Funding Guide, 2019 [61]). The author advocates for creating specific technical information necessary for long-term preservation, as well as borrowing and adapting from other disciplines. While long-term preservation and access may have been considered from the outset, the author also argues that not enough is done to establish a digital stewardship framework approach. The Digital Preservation at Oxford and Cambridge (DPOC) project (2006–2018) (Digital Preservation at Oxford and Cambridge, 2016 [21]), provides the opportunity to look more holistically at how digitised and born-digital content is created, acquired, preserved and made available.

S. Langley (✉)
Cambridge University Library, University of Cambridge, West Road, Cambridge, UK
e-mail: szl20@cam.ac.uk

H. Kremers (ed.), *Digital Cultural Heritage*,
https://doi.org/10.1007/978-3-030-15200-0_15

At Cambridge University Library (CUL), a case study approach has been adopted, in order to better understand the needs of different ‘classes’ of digital content. Examples discussed include digitised fragments from the Taylor-Schechter Cairo Genizah Collection and the interactive data in the Kymata Atlas, illustrating two very different challenges of stewarding digital content. Through the case study research, the author and colleagues have identified that digital cultural heritage and digital humanities projects often develop a website or online resource as a mechanism for providing access to digital content project outputs. If not adequate planned for, digital content is at risk of becoming inaccessible after a project ends. Migration of files and various web archiving approaches are examined as possible preservation techniques, as well as other digital capture and documentation approaches more commonly used in contemporary art, time-based media and multi-platform archiving domains (Langley et al. in Proceedings of the 19th International Symposium on Electronic Art, ISEA 2013, 2013 [43]). Considering how to preserve and provide access to digital content right from the beginning of a project is essential. Taking a holistic digital stewardship approach—while learning from the lessons of past projects and borrowing from similar disciplines—can assist in better preparing for the end of a digital cultural heritage or digital humanities project.

**Keywords** Digital stewardship · Digital preservation · Digital humanities · Digital cultural heritage · Digital Preservation at Oxford and Cambridge · DPOC · Cambridge University Library · Migration · Web archiving · Video screen capture · Data management plans · Performing arts · Technical riders · Three-Legged Stool Model for Digital Preservation · Taylor-Schechter Cairo Genizah Collection · Kymata Atlas

## 1 Introduction

Long-term management of digital content, whether from the digital cultural heritage, digital humanities or another discipline, presents a variety of challenges. As is the nature of research funding, the ability to plan for the long-term is often hampered by short-term project-based funding. Given the rapid churn of technology, inbuilt obsolescence (in devices, computer hardware and operating systems etc.) and interdependencies between software, operating systems, hardware and peripherals, keeping digital content accessible in the long-term is both non-trivial and time-consuming. Without ongoing funding resources or planning that is established early on, digital content outputs can be at the receiving end of rapid, pragmatic decisions when a project draws to a close. A website may be taken offline or left to languish, quickly become ‘technologically frozen’ until further resources can be secured [35]. Digital content that is left unmaintained is at risk of losing functionality, potentially becoming completely inaccessible. In other words, neglecting digital content will not work [17]. Online resources can also become a risk to wider technological infrastructure if they are hosted or they operate within the framework of a larger organisation. Without timely system upgrades and ongoing active management, the platforms established to provide access to digital content can become a significant security risk.

While the digital preservation discipline has been recognised internationally for over two decades,[1] for many organisations, in-depth operational digital preservation activities are only just commencing. The field of digital curation (which emerged out of eScience) appeared in the mid-2000s, in order to support reuse of digital content [28]. It was recognised that not only was the preservation of digital content necessary, the selective curation of it was also essential. While digital preservation efforts have been typically focused within the library and archive sectors, digital curation developed within the academic and research domains. While these two highly related disciplines have developed in parallel, there has not yet been enough cross-over between both disciplies. An area of work that warrants a far better collaborative approach is between born-digital archival practices and Research Data Management (RDM). At times, it is not easy to identify the difference between, or to be able to separate born-digital archival materials from research data. Indeed, relationships between individual items may mean that the digital content could be classed as either. Pathways for transfer of digital content into a collecting institution's custody may be the determining factor as to whether digital content is classified as an archival collection or as research data.

Now is the time to better conceptualise and act on the management of digital content more holistically. Considering the overall stewardship of digital content, rather than allowing digital content to be pigeonholed by one discipline or another (particularly due to the limitations of either the digital curation or digital preservation disciplines), is critical. Taking a holistic approach to preserving and managing digital content allows for mitigating issues that otherwise may not be apparent until further down the line, where future necessary interventions may prove costly, time-consuming or simply not possible. Efforts can begin by planning early on. Long-term thinking regarding digital content that may be output from a digital cultural heritage, digital humanities or other research project is essential to consider from the outset. In other words, commence planning the project's 'funeral' during its conception.

## 2 Digital Preservation at Oxford and Cambridge

The Digital Preservation at Oxford and Cambridge (DPOC) project [21]—a two-year collaboration between Cambridge University Library (CUL) and the Bodleian

[1] Digital preservation efforts at CUL began in 1998, or earlier. The Consortium of University Research Libraries (CURL) Exemplars for Digital ARchiveS (CEDARS) digital preservation project collaboration between the Universities of Cambridge, Leeds and Oxford, which commenced in April 1998, put Cambridge 'on the map' in terms of international digital preservation activities. Another organisation who was an early adopter of both digital preservation and web archiving practices was the National Library of Australia (NLA). The NLA's web archiving service, PANDORA, was launched in January 1996, while the NLA's first digital preservation policy was published in July 2001.

Libraries, Oxford—provides the opportunity to look more holistically at how digitised and born-digital content is created, acquired, preserved and made available. Based on the Three-Legged Stool Model for Digital Preservation [38], three Digital Preservation Fellows have been employed at each institution. A Policy and Planning, Outreach and Training, and Technical Fellow in each location allows for a more encompassing approach to be taken during the lifespan of the project. The DPOC project examines digital content from previous digital humanities projects as well as born-digital acquisitions and research data. While the vast majority of items being examined in this project are in digital form (with the exception of selected analogue audiovisual carriers), there may be considerably different approaches to how digital content is acquired, captured and access provided in the long-term. Amongst a range of other project deliverables, a series of case studies are being undertaken by CUL, illustrating several different approaches for acquiring, capturing and preserving digital content.

## 2.1 *Digital Content Classes*

Six ‘classes’ of digital content were identified as being held within CUL’s collections (illustrated in Table 1). These classes represent the types of digital content commonly held in academic library collections around the world. The classes were determined by the typical ‘collecting area’ functions of academic research libraries, including published digital content (Class 4), Special Collections (Classes 1 and 2), research publications and research data (Class 3) and digitised content (Class 5), including both still image and audiovisual digitisation. Commonly found in Galleries, Libraries, Archives and Museums (GLAM) sector collections is also a considerable amount of digital content that has been created internally, within an organisation (Class 6). Digital content created by an organisation is often documentation (typically photographs and/or videos) of public events, or even digital promotional materials. Class 6 content also covers documentation of physical collection items, such as photographs of a collection item undergoing conservation treatment. These images are typically created before (and sometimes after) repairs are carried out by conservators.

These six classes are used to guide the implementation of CUL’s Digital Preservation Policy [4] and other emerging digital acquisition and preservation plans.

## 2.2 *Case Studies*

As part of the DPOC project, CUL has been undertaking three separate case studies: digitised image content (Case Study 1), a born-digital acquisition (Case Study 2) and a research data project (Case Study 3). The three case studies were selected from 40

**Table 1** Classes of digital content held in Cambridge University Library's collections

| Class | Type | Description |
|---|---|---|
| 1 | Born-digital personal and corporate records | Digital archives of significant individuals or institutions |
| 2 | Born-digital university records | Selected digital records of the University of Cambridge |
| 3 | Research outputs | Research data, research publications, electronic and digitised theses, scholarly digital editions, supplementary research relating to digitised content and associated materials[a] |
| 4 | Published born-digital content | Web archives, eBooks, born-digital maps, born-digital music, digital ephemera, published born-digital content held on physical format carriers[b] and copies of electronic subscription materials (archival and/or access copies, as permitted by agreements) etc. |
| 5 | Digitised content | Digitised image content:<br>Two-dimensional (2D) photography and three-dimensional (3D) imaging etc.<br>Digitised audiovisual content: Moving image (film and video) and sound recordings etc. |
| 6 | In-house created content | Photography and videography of events and lectures, photos of conservation treatments etc. |

[a] Associated materials are considered to be data that provides context and assists in the interpretation of digital content. It may also refer to files that are essential for rendering digital content. Associated materials may include algorithms, code, diagrams, documentation, sidecar files, scripts, transcripts etc.

[b] Physical format carriers include magnetic tape (carrying analogue or digital audio and/or video content), motion picture film (carrying optical moving image and may include audio content), disks (zip disks, $3\frac{1}{2}$ inch and $5\frac{1}{4}$ inch floppy disks, carrying data), optical media (such as Compact Discs, Digital Versatile Discs and Blu-ray discs, carrying data, audiovisual or multimedia content) portable hard disk drives or USB flash drives (carrying data)

potential candidates, that represented a broad range of digital collection items, from each of the six digital content classes.

It was essential that each case study represented a different class of digital content. In addition, the case studies were chosen based on a range of criteria including: complexity, frequency and/or volume, significance, urgency, uniqueness and value to end users and/or stakeholders [40]. Each case study was ranked against seven key 'stages' that digital content passes through when being transferred to a collecting institution, and as it is managed and preserved. The stages selected included: 'Appraise', ‘Acquire’, ‘Pre-Ingest’, ‘Ingest’, ‘Store & Manage’, ‘Preserve’[2] and ‘Deliver and/or Provide Access’. These stages were loosely based on the Digital Curation Centre’s Curation Lifecycle Model [18] and the author’s previous experience in handling born-digital content.[3] Given the strong driver to address digital content holistically, as part of Case Study 2, the Digital Stewardship End-to-End Workflow Model was developed. This workflow model outlines a total of 14 different stages—from conceptualisation through to use and reuse [41]. Coupled with the Digital Stewardship End-to-End Workflow Model, the Digital Streams Matrix [42] was developed by the author, to aid decision-making when identifying and selecting actions and processes needing to be carried out on the different classes of digital content. For example, the Acquire stage of the Digital Streams Matrix outlines several ways of capturing web-based content.

Many digital cultural heritage and digital humanities projects provide their project outputs online. Two of the three CUL case studies (Case Studies 2 and 3) contained websites. In their simplest form, these websites may be the only location for digitised images created as part of a digital humanities project.[4] While web archiving was not the main focus of the case study work, some initial explorations into how these websites—which were critical to both Case Study 2 and 3—could be archived, was attempted. The approaches investigated are based on several common use cases found as a result of surveying CUL’s digital collections. During the Bodleian Libraries collection surveying activities, DPOC counterparts at Oxford discovered that many projects containing digital content were also created in the online environment, often as websites. Based on a number of identified risks, Bodleian Libraries concluded that the digital content held in these websites would need to be ‘rehoused’ in the near-future. As a result of exploring the various approaches for small-scale capture of websites (as part of a larger digital acquisition process), a selection of available tools were identified, suitable for use in a range of different scenarios. These tools and approaches are discussed further on in this paper.

[2] While digital preservation considerations should be addressed at each stage, typically ‘preserve’ in this sense refers to in-depth digital preservation activities, such as ‘preservation actions’. A preservation action may consist of a number of tasks, from a simple checksum verification, through to migrating a batch of files from one file format to another.

[3] Previous experience handling born-digital acquisitions was obtained through acquiring, managing and preserving born-digital collections while employed at several significant Australian cultural institutions including the NLA, the National Film and Sound Archive of Australia and the State Library of New South Wales.

[4] As part of the DPOC, the Digital Preservation Fellows at the Bodleian Libraries, Oxford found in excess of 40 websites at the University of Oxford, containing mainly digitised image outputs. Bodleian Libraries are in the process of consolidation; extracting digital content from identified websites and storing the files in one of the two Bodleian Libraries’ digital repositories: Digital.Bodleian or the Oxford University Research Archive (ORA).

## 3 Digital Stewardship

Digital heritage, as defined by the United Nations Educational, Scientific and Cultural Organization (UNESCO), encompasses digital content from a range of 'different communities, industries, sectors and regions' [59]. When digital content is transferred into the custody of a collecting institution, it is managed alongside other digital content from a vast array of professional fields. In order to retain the original meaning and intent of the digital content, context is critical. For digital content to be adequately archived and preserved in the long-term—particularly if custody is transferred to a collecting institution—comprehensively organising and maintaining digital content should not be the final step taken by a content creator or producer, just prior to transfer. Steps towards preservation should have already taken place during the digital content's lifespan. Digital preservation must be 'baked in' throughout the lifecycle of the digital content [58]. As active management should occur right from the start, 'stewardship' is a better way to conceive of managing digital content. For this reason, it would be wise for any digital cultural heritage or digital humanities project to adopt a 'digital stewardship' approach.

Digital stewardship brings together the concepts of both digital preservation and digital curation. In 2010, the National Digital Stewardship Alliance (NDSA) [49] was launched, as an initiative of the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP) [46]. National Digital Stewardship Residency recipient Jaime McCurry, defined digital stewardship as encompassing [45]: "…all activities related to the care and management of digital objects over time. Proper digital stewardship addresses all phases of the digital object lifecycle: from digital asset conception, creation, appraisal, description, and preservation, to accessibility, reuse, and beyond." [48]. Digital stewardship provides a framework for long-term thinking to ensure that preserving and managing digital content for the long-term is not merely an afterthought.

## 4 Digital Curation and Preservation Models

There are a wide variety of models for conceptualising and guiding the management of digital content. Three models considered useful to the fields of digital cultural heritage and the digital humanities are briefly discussed.[5]

The Three-Legged Stool Model for Digital Preservation [38] considers the 'Organizational Infrastructure' (the 'what'), the 'Technological Infrastructure' (the 'how') and the 'Resources Framework' (the 'how much') as equally important parts necessary for undertaking digital preservation activities. It is common to consider digital preservation only as a technical problem with a technological solution. The

[5] Two of these models have already been referenced in the paper thus far. The author's newly developed alpha release Digital Stewardship End-to-End Workflow Model has been mentioned in the context of CUL's Case Study 2. It is not discussed further in this paper.

Three-Legged Stool Model tries to dispel this myth; identifying that the Technological Infrastructure—the 'equipment, software, hardware, a secure environment, and skills to establish and maintain the digital preservation program'—is just one of three essential aspects. The 'Organizational Infrastructure' includes 'policies, procedures, practices and people' and the 'Resources Framework' that encompasses 'start-up, ongoing, and contingency funding' are equally critical. Without addressing each of the three areas, successful digital preservation efforts are not possible. For digital cultural heritage and digital humanities projects, identifying and addressing the 'what', 'how' and 'how much' right from the start, will assist in improved sustainability of digital content in the long-term.

The Digital Curation Centre's Curation Lifecycle Model [18], first published in draft in 2007 [29], comprises several layers of 'actions': Full Lifecycle Actions,[6] Sequential Actions and Occasional Actions. The model is most useful for its eight Sequential Actions: 'Conceptualise', 'Create or Receive', 'Appraise and Select', 'Ingest', 'Preservation Action', 'Store', 'Access, Use and Reuse' and 'Transform'. These are the sequential actions that digital content typically passes through, with the latter seven of these actions being cyclical. Digital content can continue to be managed via this sequence of actions throughout its entire lifecycle. The Occasional Actions of 'Reappraise', 'Migrate' and 'Dispose' may become relevant to some digital content, depending on the circumstances. This decade-old model continues to be enhanced, as the field of digital curation matures and evolves.

The CLOCKSS [12] Threats Model and Mitigation Strategy [13] is another holistic (risk managed) approach to digital preservation. The model outlines various disasters, errors, failures, and obsolescence that could occur, which would put digital content at risk. These include commonly thought of failures of media, hardware, software and network services, as well as economic and organisational failure. Obsolescence (of media, hardware and software) can also play a part in placing digital content at risk. Communication and operator errors are considered a risk, as are natural disasters, internal and external attacks. All are factors that should be risk managed. This Threats Model underlies the Lots of Copies Keeps Stuff Safe (LOCKSS) [47] technology, and was formalised in 2005 as a 'bottom-up' approach to developing digital preservation system requirements [55]. Considering, at least briefly, each of these risks at the start of a digital cultural heritage or digital humanities project will better equip the project to manage problems as they arise. This would also aid in planning for long-term preservation of, and access to, any digital content created. Not only relevant to digital cultural heritage and digital humanities researchers, these models should also be considered by digital content creators and producers, as well as staff working in collecting institutions.

[6] The Full Lifecycle Actions are higher-level considerations centred around 'Description and Representation Information', 'Preservation Planning', 'Community Watch and Participation', and 'Curate and Preserve'. For reasons of brevity, they are not discussed in this paper.

## 5 Migration

For the past decade, CUL has been digitising items in its collections and making these images available via the Cambridge University Digital Library (CUDL) [5]; a delivery system with an embedded digital image viewer and other functionality.[7] While the focus has been on the digitisation process and provision of access, CUL has realised that digital image files need to be preserved in sustainable ways, hence the digital preservation scoping work being undertaken as part of the DPOC project. The DPOC project has drawn attention to the fact that CUL's digitised image collections are held in multiple locations. This means that not all digitised images are managed in Extensis Portfolio, the current Digital Asset Management System (DAMS) in use at CUL.

For smaller digital cultural heritage or digital humanities projects where a DAMS is not available, it is more likely that the output of digital content has been published via an online mechanism or resource. Digital content that is made available via a website (where the website is built utilising a back-end database), may ultimately be easier to migrate into a digital preservation system or digital repository, than digital content embedded into individual webpages. With the vast array of web technologies used to create a website, managing a migration process must ultimately be handled on a case-by-case basis. (It should be noted that for digital content embedded in a website that does not have an underlying database, 'screen scraping' may be the only option.)

For digital content held in a database of some form, migration is usually the best approach. Migration refers to the process whereby digital content (and associated metadata) is transferred—typically in large 'batches'—from one environment, system or repository (such as database or website) to another (such as a DAMS or digital preservation system). Several basic stages as part of a migration process include analysis, preparation, assessing and undertaking the migration [63]. Any migration process must be well planned. Some necessary steps include (but are not limited to):

- Analyse the digital content that is held in the environment, system, repository or location you wish to migrate from[8];

[7] The CUDL provides zooming functionality to view digital images, using OpenSeadragon technology. The delivery system is also going through a redevelopment to improve functionality for digital images—such as implementing the International Image Interoperability Framework (IIIF)—and to provide mechanisms for accessing digital content in other formats. The development of the CUDL was generously funded by The Polonsky Foundation and the Andrew W. Mellon Foundation.

[8] For proper analysis of digital content prior to a migration process, identifying the variety of files is essential. Running digital preservation software tools such as DROID (which utilises The National Archives of the UK's PRONOM file format registry), Siegfried, the Open Preservation Foundation's JHOVE, or FITS across all of the digital content will assist in identifying the formats of the files, and potentially also characterise and validate the files (depending on which tools are used). Undertaking this type of in-depth analysis will pre-empt ingest issues or failures as part of a migration process, particularly when migrating digital content into a digital preservation system.

- Analyse the associated metadata of each file (including for 'complex digital objects',[9] where it is necessary to identify the metadata critical for retaining the integrity of the entire 'digital object');
- Define both mandatory and optional metadata (and identify what metadata is unnecessary to migrate)[10];
- Document all known 'edge cases'[11] (for both digital content and metadata), and test the ability to export the digital content and metadata that falls within these edge cases;
- Define what metadata does not exist and must be created as part of a migration process (e.g. preservation metadata, including 'agent information'[12]);
- Generate checksums for each file[13];
- Create a technical manifest[14];
- Develop a metadata crosswalk to map from the 'old' environment, system, repository or location being migrated out of, to the 'new' environment, system, repository or location being migrated into—using standards-based metadata schemas[15];
- 'Data washing' of any metadata that doesn't correctly map to the new metadata schema and/or environment, system, repository or location[16];

[9] A complex digital object is a type of digital content that consists of a number of files (often of different formats) that make up the components of a single digital item. eBooks, websites and interactive multimedia works are considered some of the types of complex digital objects.

[10] If in any doubt, retain metadata rather than discard it. This is particularly important if analysis is not comprehensive, or if the function of certain metadata isn't fully known or understood.

[11] An 'edge case' is a less common situation that either does not follow the expected rules within a system or falls at the extreme boundaries of the system. Any systems that have used a 'Band Aid' approach to managing digital content or metadata can be expected to contain edge cases. Without analysing and testing for edge cases, it is likely that some digital content and/or metadata will be lost as part of a migration process.

[12] Information about 'Agents'—as defined in this instance by the PREMIS preservation metadata standard—would refer to information about the software tools and version(s) of the tools used to create or make changes to files.

[13] A checksum is a cryptographic hash, which is typically represented as an alphanumeric code that is unique to each file (with the exception of collisions). There are a range of different checksum algorithms, such as SHA-1, SHA256, SHA512, MD5 etc., that are often used. Once a checksum has been created for a file, this alphanumeric code can be used to verify whether a file has been modified. For as long as the alphanumeric code remains the same, this indicates that the contents of a file has not been altered.

[14] A technical manifest is a record of a group of files. Typically this would be used as information about a 'complex digital object', digital collection, or batch of files (as part of a migration process). As a minimum, a technical manifest should include for each file, the checksum, and the corresponding filename and file path.

[15] Metadata standards that may be useful for developing crosswalks, are highly dependent on the system(s) being migrated from and to. These must be developed on a case-by-case basis, and are also influenced by how metadata has been recorded in a database, system or repository. Common metadata standards that may form part of a crosswalk include Dublin Core, EAD, EBUCore, METS, PREMIS, TEI etc., depending on the type of digital content and systems used.

[16] Also known as 'data cleansing' or 'data cleaning'. This may include removing duplicate data or adding in specific metadata that does not already exist within the current database, repository or system.

- Test export of metadata—including metadata in relation to digital content (so relationships between each file and its associated metadata is retained);
- Test export of digital content to ensure the quality of the digital content (including any metadata embedded in files) is maintained during the export process;
- Establish an in-depth understanding of the environment, system, repository or location the digital content is being migrated to (including which metadata standards and file formats it supports);
- Configure the new environment, system, repository or location, including ingest workflows (and other workflows that are supported);
- Fully test ingest workflows in the new system to ensure that content batch-migrated from the old environment, system, repository or location can be successfully ingested.[17]

While similar tasks for each migration are likely to apply, each collection of digital content consists of different elements and may be arranged in vastly different ways. Therefore, each migration process will differ. The order of tasks may switch around and not every step may be necessary. Alternatively, for environments, systems or repositories that have not been used in standardised ways (or importantly, how they have been used has changed over time), additional steps may be required. Hence, analysis of the digital content, metadata and both the old and the new environment, systems, repositories or locations are necessary. A lesser degree of analysis may be acceptable if a comprehensive understanding and documentation already exists.

Once the outlined steps (identified above) and any additional requirements have been determined (and documented), a thorough understanding of the digital content and metadata will have been achieved. It is at this point that developing a migration strategy is possible. The migration strategy should be accompanied by quality control testing and error handling procedures (as issues will arise). Additionally, consider the migration process as not only an export from one system to another, but also an opportunity to improve metadata along the way. While the focus of this exercise may be migration, it is essential to develop an in-depth understanding of how newly created digital content (and its associated metadata) is ingested into the new system, using the new workflows. While this may seem like a considerable amount of work, cutting corners in the analysis and preparation stages may actually cost time and/or resources if issues occur, and have not been adequately planned for.

### *5.1 The Taylor-Schechter Cairo Genizah Collection*

CUL's Taylor-Schechter Cairo Genizah Collection is considered 'the world's largest and most important single collection of medieval Jewish manuscripts' [6]. The almost 193,000 fragments held at CUL were discovered in the Ben Ezra Synagogue in Egypt,

[17] Testing is necessary to ensure that there are no 'show stoppers' when a full migration process takes place. Not undertaking adequate testing may mean that the migration process fails, takes much longer and/or the costs of the migration work increase.

documenting approximately 1000 continuous years of Jewish history in the Middle East and North Africa. The ability to share these digitised fragments with the world, via the CUDL [5], is one of CUL's successes.

CUL's digital image content is stored in several locations. The diversity of storage locations is often dependent on how the digitisation project was funded, or indeed, the department that undertook the digitisation work. While digital image preservation master files are stored in networked file stores, other projects (related to both digitised and born-digital content) have resulted in preservation master files stored in more than one instance of a DSpace repository instance, such as the Taylor-Schechter DSpace.[18] In order to support the Taylor-Schechter DSpace, at the time the DSpace instance was created (in 2009), an 'Operator Manual' was also developed. The Operator Manual includes details on filenaming protocols, workflows and backups. Unfortunately, over time, this manual has not been updated, and so no longer fully reflects the Taylor-Schechter DSpace instance.

As part of the DPOC project, preservation master files (high-resolution images) of fragments from the Taylor-Schechter Cairo Genizah Collection—that are stored in the Taylor-Schechter DSpace—have been identified as being 'at risk'. This is due to the fact that these images have not been actively managed since the completion of the digitisation project. Without project funding, there was no dedicated person or team assigned to actively manage the preservation master files, as CUL do not have a dedicated digital preservation staff member or program at present. In addition, the Taylor-Schechter DSpace has not received scheduled updates of the repository software, and adding new users to access the Taylor-Schechter DSpace content is no longer possible. (This means that only a select number of CUL staff continue to maintain access to this repository.) A variety of issues with the Taylor-Schechter DSpace have manifested over the past decade, resulting in the need to decommission this legacy system and migrate the preservation master files into a digital preservation system.[19] As part of the DPOC project, some analysis has taken place, identifying a range of issues. For example, original filename identifiers[20] are only stored as metadata due to the repository software auto-renaming files on ingest. Additionally, selected preservation master images from the Taylor-Schechter DSpace have been individually migrated to a networked store, when an image is ordered by an external (or internal) client. A brief plan of action to export over 360,000 preservation master files from the Taylor-Schechter DSpace has been drafted. However, this is only one component of a full migration strategy. While there may be a variety of technical challenges to overcome, the main challenge is not technical. It is a resourcing issue. Given the shared responsibilities of this digital content, which currently lies between

[18] This is one of several instances of an internal 'dark' DSpace repository at CUL. CUL also provides the University of Cambridge's Apollo Open Access Repository, which is a publicly available DSpace instance [60].

[19] At the time of authoring this paper, CUL is working towards a business case for the implementation of a digital preservation system.

[20] CUL has developed its file naming 'protocols' based on the Classmark scheme used for physical items in CUL's collections [7]. At present, CUL does not have an organisation-wide Persistent Identification (PI) scheme for file naming.

CUL's Digital Content Unit (DCU), Digital Library Unit (DLU), Digital Initiatives and Strategy (DIS) and the Taylor-Schechter Genizah Research Unit, a small project team must be put together to develop a migration strategy. There are likely to be a range of edge cases and other complexities that must be factored in, requiring input from subject matter experts, digitisation experts and IT staff.

Some lessons can already be garnered from this particular example. Standardisation is essential, and supporting documentation must be created and then adequately resourced, in order to maintain the documentation and keep it up-to-date. Standardisation does not only refer to the creation of the digital content, it also refers to the way in which supporting documentation is created. Rather than 'operator manuals' or 'protocols', a suite of standardised documents including policy, standards, procedures and guidelines (PSPG) as well as strategies and plans are necessary. For digital cultural heritage and digital humanities projects, also establishing standardised documentation right from the start will mean an easier process of managing digital content throughout the lifetime of the project. Standardised documentation will also aid in the long-term, particularly when transferring digital content into the custody of another organisation.

## 6 Bundling and Bitstream Preservation

One of the simplest ways of archiving digital content—particularly a set of files—is to zip[21] them up into a single 'bundle'. Bundling a set of files ensures that they remain together. This should be coupled with a checksum. Generating a checksum hash and then validating these hashes at regular intervals is the method used to monitor changes to digital content. A wide range of checksum verification tools are available, which range from simple Command-Line Interface (CLI) to Graphical User Interface (GUI) tools with additional features.[22] A single checksum can be generated for the bundle, rather than a checksum generated for each file within the bundle. Once the zip file and its associated checksum have been created, ensuring the 'fixity'[23] of this bundle can be a somewhat trivial exercise. While this may be the simplest approach, this does not provide for any sophisticated methods of managing the digital content, should

[21] A zip file is often a single file that represents a folder (or directory) of a set of files that have been compressed together. Various types of compression can be applied to make the size of the zip file smaller for the purposes of transfer or storage. Compression does not always have to be applied. This paper does not go into the considerations and technicalities of applying compression to zip files and how this may affect preservation of digital content.

[22] A range of tools for generating and verifying checksums are available. These include (but are not limited to) Bagger, bagit-python, checksum+, Exactly, hashcheck, Hashdeep, Fixity and ExactFile. Some are CLI tools, while others are GUI software. It should be kept in mind that specific tools may only be available for particular operating systems and computer platforms.

[23] 'Fixity' is the measure used by the digital preservation discipline to ensure that no unauthorised change to digital content occurs. Fixity refers to the checksum(s), filename and file path—as a minimum—that are generated and/or recorded for a specific file.

a bundle somehow be modified. A checksum for each file in a set of files would provide further granularity and may make troubleshooting easier in the long run. Hence, when a zip bundle is created, sometimes a technical manifest containing the fixity of each of the files in the set is also generated. This is an additional safeguard, should it be needed for any future monitoring or troubleshooting.

The notion of bitstream preservation is to ensure that at least one copy of the data (0s and 1s) is maintained at all times, and to ensure the integrity of this data (the 'bitstream'). Maintaining the fixity of digital content is a core concept of digital preservation, and ensuring the integrity of the data is typically done by 'fixity checking'. The typical method used for fixity checking involves regularly verifying that checksum hashes for each file remain unchanged [22]. Depending on the tool used to generate a checksum hash, unless there is a comprehensive understanding of the formatting of the fixity information, it may be wise to use the same tool to undertake fixity checking (particularly where a researcher has a lower digital or technical literacy). Documentary evidence to prove these requirements have been met is also necessary [1].

Fixity checking by way of monitoring checksums does not prevent intentional or unintentional changes from occurring. It is only a method for detecting change. Whether modification of a file is intentional (potentially malicious) or unintentional, or if the file has been corrupted, this is unable to be determined by fixity checking. While many other approaches to archiving digital content described in this paper require certain infrastructure, fixity checking is the simplest method for managing digital content. This is particularly the case for individual researchers, where institutional data management and storage is not available.

## 7 Packing Down and Putting on Ice

If a digital cultural heritage or digital humanities project faces a temporary resourcing challenge, one method of archiving the digital content (particularly if there is a good chance the project may recommence at some point in the future point) may be to pack the project down, in a way where it is effectively being 'put on ice'. This could be thought of as the 'cryogenically frozen' approach. At some point in the future—when the necessary resources have been secured—it can be 'brought back to life'. Developing a thorough content model and/or data model at an early stage of a project is crucial in order to utilise this approach.

While somewhat embryonic, one example of this approach is the Australian project, Ozmeka [51]. Based on Omeka [50], the Ozmeka version (of the Omeka code) has been developed with the intention of supporting eResearch projects. It is an attempt at a standardised way of handling the digital outputs of research projects, where they have been presented online. This approach is being trialled with the eResearch project Dharmae, the Data Hub of Australian Research on Marine and Aquatic Ecocultures [14]. It is hoped that over time, being able to 'cryogenically freeze' and bring back to life digital projects will be less cumbersome. At present, one of the challenges faced by the RDM community is the grey area between ‘active research data’ and ‘archival research data’. Digital content can easily switch between being ‘active’ and needing to be preserved for the long-term. However, the infrastructure and mechanisms available to move digital content between these two states (and often, two different storage environments) is not easily facilitated. Far better methods for allowing the ‘spin up’ and ‘spin down’ of digital project environments are needed.

### *7.1 Databases and Code Repositories*

Where archiving and preservation of a database is important—not only preserving the digital content and/or the metadata held in a relational database (including web-based databases), but the database itself—then this can be achieved by using the SIARD (Software Independent Archiving of Relational Databases) format [56] and SIARD Suite [57]. For digital cultural heritage and digital humanities projects that are generating a large amount of data, where the relationships between the data elements are of considerable importance, preserving the whole database may be an approach to consider. Due to the finite nature of the DPOC project, preservation of databases was not considered a priority and so this option was not investigated further.

For code-based projects, a ‘good practice’ approach should be taken when managing, versioning and sharing code. Where hosting arrangements are part of a digital cultural heritage or digital humanities project—such as working in an academic or another institutional environment—other factors may need to be taken into account. For example, this may include committing code to a local code repositories, rather than using GitHub [26]. Where code is stored, it must be documented. Documentation should include information on dependencies between other aspects of a project, as well as instructions on how to install, configure and run the code. Ideally documentation should be standardised as PSPG. When working in the context of an academic or other institutional environment, be pragmatic and utilise any existing PSPG templates.

## 8 Alternative Capture Methods

Context is crucially important for digital content. Where the complexity of the digital content doesn’t easily facilitate more straightforward digital archiving approaches (such as those that have already been discussed in this paper), documentation methods may be suitable. These documentation methods can help in providing an understanding of how the digital content was used, and are particularly useful for use with interactive digital content.

### 8.1 Video Screen Capture

A simple time-based method to document a 'single pass' through a complex digital work—whether this is a website, game or an interactive experience—is to record a Video Screen Capture (VSC).[24] A VSC records the image content that appears onscreen (plus associated audio), and produces a single channel video file. If attempting to document an interactive environment (containing an infinite number of pathways and experiences), then it is recommended that several VSCs are recorded. Ideally, each VSC would document a different 'journey' through the interactive environment. Capturing several different journeys—from different perspectives—-may be a suitable approach for documenting a complex environment. This digital archiving approach is already in use for capturing multi-platform content experiences [43]. Together with the original files (such as project files) plus other documentation, VSCs may provide researchers of the future with a better understanding of how the digital content operated, and what the experience was like for a user.

### 8.2 The Kymata Atlas

An example of where using the VSC approach was considered suitable, was for documenting the user interactivity of the 'Surface Viewer' interface of the Kymata Atlas [39], illustrated in Fig. 1. The Kymata Atlas—a research resource from the MRC Cognition and Brain Sciences Unit at the University of Cambridge—provides a partial set of information processing pathways mapping of the human brain (represented as 'functions'). The Surface Viewer allows for interactive visualisation of the datasets contained within this research resource. Users can move their mouse over the various sectors of the brain, which displays information found at various coordinates.

Using VSC provides a simple mechanism for archiving the interactive interface, and displaying how the Surface Viewer could be used. However, as the Kymata Atlas is a significant and internationally-relied upon resource containing numerous datasets, other approaches would also be required in order to fully archive the entirety of this online resource.

### 8.3 Documenting in Context

A more advanced approach to archiving websites via VSC is illustrated in Robert Sakrowski and Constant Dullart's netart.database [23]. In addition to recording a VSC, a camera (mounted behind and to one side of the user) records an individual's interaction with a website, and the computer hardware the website is displayed on.

[24] In the Mac environment, QuickTime (version 10.4) is capable of creating a VSC.

Capturing this additional element provides another layer of information for future researchers. The netart.database project focusses on a particular sub-set of media art culture; that of net.art. This additional documentation will allow future researchers to observe how interactive net.art works were experienced, showing a user engaging with the computer hardware and browser software available at the time.

While the netart.database may not be a preferred method for documenting some digital cultural heritage or digital humanities projects, for Human-Computer Interaction (HCI) and Information Communications Technology (ICT) researchers, this approach may provide a simple and straightforward method for capturing the interaction between humans and computers (and/or other devices). The beauty of this method is—compared to motion capture[25] approaches—its simplicity and low-cost.

## 8.4 *Rhizome's Webrecorder*

Developed for Rhizome, the US media arts organisation (and initially funded by the Andrew W. Mellon Foundation), the Webrecorder tool [54] is another time-based approach to archiving websites. The open-source project allows for the 'recording' of websites, with the output produced as WARC (Web ARChive) files.[26] A user

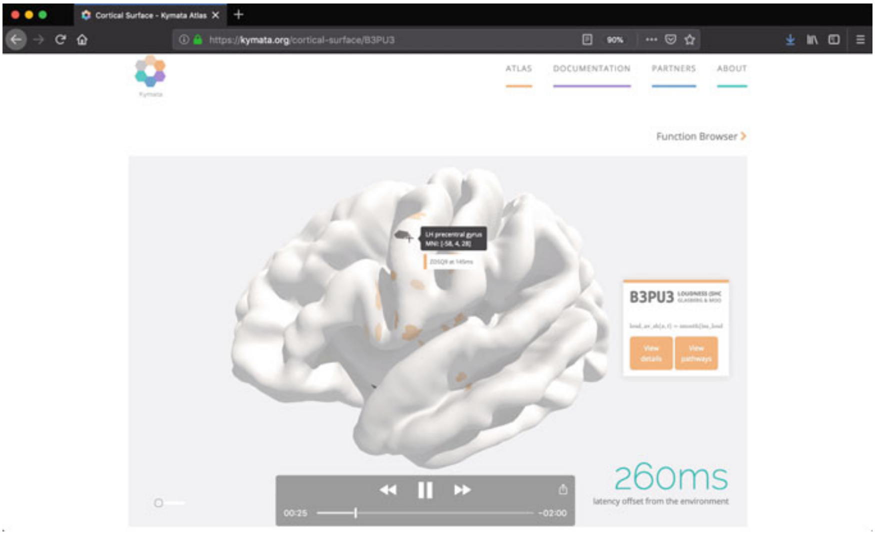


Fig. 1 Still from a Video Screen Capture of the Kymata Atlas Surface Viewer

[25] Motion capture is an approach for capturing the movement of humans or other objects in a three-dimensional space, typically producing a dataset that represents positions in space over a certain time period.

[26] A revised version of the Internet Archive's ARC file format.

wanting to capture a website enters the URL and navigates around the site. All actions are automatically recorded. When the required interaction with the website is complete, all the user needs to do is to stop the ‘capture’ button. This ‘session’ is stored for a limited period of time, allowing a user to decide whether to turn the recorded session into a permanent web archive file. Like other web archives stored using the WARC file format, in order to view the contents of the WARC files, ‘playback’ tools are required.The International Internet Preservation Consortium’s (IIPC) OpenWayback software [31] is one tool that can be used for this purpose.

## 9 Hosted Infrastructure Arrangements

For some digital cultural heritage or digital humanities projects, a hosting arrangement entered into with an organisation may be the best and most cost-effective approach. Organisations, rather than individual research projects may have at their disposal a range of infrastructure that simply cannot be sourced for a single project alone. That said, transferring the hosting of a complex website or online resource to another organisation may be time-consuming. Additionally, some hosting organisations may not support the software the online resource has been developed with, the underlying technical infrastructure that the online resource is dependent on, or the access methods provided to the digital content.

If a hosting arrangement is one of the longer-term possibilities being considered for extending the lifespan of a digital project, discussions with potential hosting organisations should begin as early as possible. Covering the costs of ongoing support and future maintenance—such as system upgrades (including supporting any ICT security compliance)—should be factored into any research grant funding applications, where permitted. If a hosting arrangement is negotiated, establishing a contract or a Memorandum of Understanding (MOU) is considered essential.

### *9.1 Design & Art Australia Online*

One example of a digital humanities project that has established a hosting arrangement is Design & Art Australia Online (DAAO). The DAAO is a collaborative scholarly eResearch tool: an online database providing access to biographical information about Australian visual artists, designers and architects etc. [15]. This project has been running for over a decade, with primary funding from the Australian Research Council (ARC), stretching across a number of different grants. The DAAO—a national partnership of art galleries, libraries and universities—is hosted by the University of New South Wales (UNSW) Library. DAAO is subject to stringent technical and security compliance, with both UNSW ICT and UNSW Library technology infrastructure.

### 9.2 King's Digital Lab

At King's College London (KCL), the King's Digital Lab (KDL) [36] provides infrastructure for eResearch and digital humanities projects to ensure the foundation of a digital project is established and stable from the get-go. These hosting arrangements allow KDL to spin-up virtual servers for various digital projects as required [37]. The added benefit of these hosting arrangements is that time and research funding is not wasted on setting up and configuring the technical infrastructure and framework for the online resource. This frees up both funding and project time, allowing for efforts to be better directed towards undertaking research activities. This method also mitigates against potential risks regarding the longer-term sustainability of digital content created as part of a project, due to relying on already established and managed technical infrastructure. Pathways are starting to be established between KDL and the KCL Library, in order to transfer digital content to library repositories (where it meets library collection development policies), at the end of the life of a digital humanities project.

### 9.3 The Casebooks Project

The University of Cambridge has been home to a number of digital humanities projects. One example is the Casebooks Project [11], a collaboration with the Bodleian Libraries at the University of Oxford. The Casebooks Project combines the fields of medicine and astrology, making almost 50,000 'cases' available online. In recent years, the University of Cambridge and CUL has been expanding their digital humanities capabilities [3, 8]. Typically, CUL does not host digital project websites, as access to digital content is provided via the CUDL. Indeed, even the Forman and Napier Casebooks [9] (as part of the Casebooks Project) can be already accessed through the CUDL. However, providing access to the full set of Casebooks via the CUDL is problematic at present. This is due to the fact that there is no 'one-to-one' relationship between a 'case' and a single page (or set of pages). A 'case' may commence midway through a page and run across multiple pages. As an interim arrangement—until such times as other viewer mechanisms are developed—the Casebooks Project is currently being hosted by CUL, having been migrated from its previous location on the Department of History and Philosophy of Science servers. Conceptualising a case is somewhat challenged by the current capabilities of typical image delivery systems and viewers, including the CUDL image viewer. Given the burgeoning of digital humanities projects and/or alternative viewing mechanisms (including other ways of representing and interacting with digital content), other options may become possible in the future.

## 10 Web Archiving Approaches

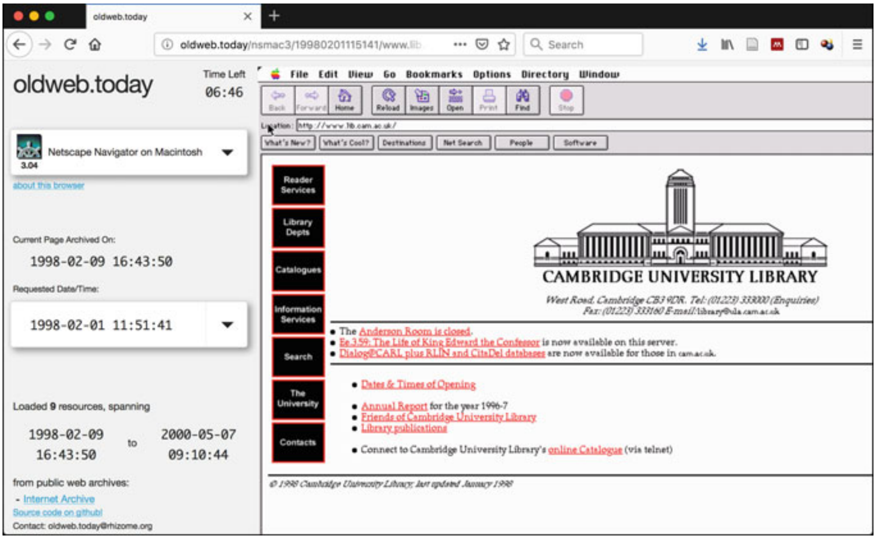
There are several freely available tools that can be used for web archiving activities, in addition to the Rhizome's Webrecorder [54]. Some tools, such as Heritrix [33], have larger-scale infrastructure requirements and demand skills that include the ability to set-up, configure and run CLI web crawlers. Other options, such as the Internet Archive's Archive-It web archiving service [32], is provided as an online service and resource. There are over 400 organisations using Archive-It. However, use of this service is restricted to established institutions such as libraries and archives, non-governmental organisations, and historical societies.

Another web archiving tool, HTTrack [30], provides both CLI and GUI options. The HTTrack tool downloads all the HTML and other static content found in a web directory and copies these files—in the same file structure—to another location. While HTTrack can be used by an individual researcher, it is not necessarily successful at capturing all digital content, particularly from dynamic websites. For example, it may miss certain types of digital content, such as audiovisual content, that is only linked (not embedded) within a website. In order to capture linked audiovisual files, these may need to be acquired via alternative methods. One method that has been used is to download the linked video files—using the Video DownloadHelper browser plugin [62]—and manually link these video files into the website files downloaded using HTTrack. This can be an onerous process if more than a handful of files need to be added back in. It is also not an exact replication of the relationship between two or more websites, and may therefore not be considered an authentic archival representation.

Capturing websites is only one part of the web archiving process. Providing access to archived websites is another piece of work entirely. Many tools used in web archiving produce files in the WARC format.[27] For websites that have already been archived into the Internet Archive's Wayback Machine [34], other playback tools are available for enabling users to interact with archived website content. For example, Oldweb.today [53] (also developed by Rhizome), provides an emulation facility, allowing users to see and experience what a website originally looked like, through various legacy web browsers (and browser versions).

For example, using the emulated Netscape 3.04 web browser, which was released in 1997, it is possible to view an instance of the Cambridge University Library website that was captured on the 9th February 1998, illustrated in Fig. 2. Providing todays' users with the opportunity to interact with websites from previous decades—using the browser software that was available at that time—allows for a more comprehensive engagement, as well as contextualisation of the archived web resources.

[27] Due to the focus of the DPOC research (and this paper), tools for the presentation of websites that have been archived in the WARC file format have not been investigated.



oldweb.today/nsmac3/19980201115141/www.lib

Time Left 06:46

Netscape Navigator on Macintosh 3.04

Current Page Archived On: 1998-02-09 16:43:50

Requested Date/Time: 1998-02-01 11:51:41

Loaded 9 resources, spanning 1998-02-09 16:43:50 to 2000-05-07 09:10:44

from public web archives:
- Internet Archive

Contact: oldweb.today@rhizome.org

Location: http://www.lib.cam.ac.uk/

CAMBRIDGE UNIVERSITY LIBRARY

West Road, Cambridge CB3 9DR. Tel: (01223) 333000 (Enquiries)
Fax: (01223) 333160 E-mail: library@ula.cam.ac.uk

- The Anderson Room is closed.
- Ee.3.59: The Life of King Edward the Confessor is now available on this server.
- Dialog@CARL plus RLIN and CitaDel databases are now available for those in cam.ac.uk.

- Dates & Times of Opening
- Annual Report for the year 1996-7
- Friends of Cambridge University Library
- Library publications
- Connect to Cambridge University Library's online Catalogue (via telnet)

© 1998 Cambridge University Library; last updated January 1998

**Fig. 2** Cambridge University Library website, captured on 9th February 1998, viewed through Oldweb.today, using the emulated Netscape 3.04 browser

## 11 Improving the Use of Data Management Plans

Research funding bodies are increasingly stipulating that for the purposes of reproducibility, datasets (that support research publications) should be submitted to appropriate open access repositories, typically those provided by the researcher's institution, or alternatively discipline-specific repositories. Researchers and staff at the University of Cambridge are strongly encouraged to submit their research publications and any associated datasets into the University of Cambridge's Apollo Open Access Repository [60]. While practices surrounding Data Management Plans (DMPs) differs from country to country, in the UK, DMPs are only created as part of the funding application or research grant proposal, and are not required to be submitted into a repository alongside research datasets. As DMPs are developed prior to research commencing, the proposed technological frameworks, software etc., may be a far cry from what is actually implemented. As a result, the initial DMP (often produced years earlier) rarely reflects the final digital outcomes and environments, developed and implemented as part of a digital project.

Taking a digital stewardship approach, DMPs could be harnessed to support digital preservation activities in the longer-term, including forming the basis for a preservation plan. Given the increasing complexity of research data, the importance of DMPs (or a similar record of the technologies used in a project including hardware, software, versions, peripherals, dependencies, standards plus other technical complexities etc.), will become more relevant over time. While it is ideal that there is involvement from subject matter experts, when trialling or undertaking digital preservation activities or actions (particularly when handling complex digital content), the reality is that this is unlikely to occur.

The following recommendations have been developed with the intent of making better use of DMPs for preservation purposes. These include (but are not limited to):

- Consider the DMP as a 'live' document—update it throughout the course of the project (preferably versioning the document);
- Submit the DMP alongside any datasets or digital content being placed in a digital repository;
- Include data citation information;
- Author the DMP for submission in a way that is usable for long-term preservation purposes—including using clear headings (and sub-headings), and write any technical specifications as lists (rather than prose);
- Include a technical questionnaire, detailing the technical specifications (encompassing all aspects of the project)—incorporate information on the code developed, any standards used, software and software versions required to use and/or reuse the digital content, hardware requirements, peripherals and any other dependencies etc.;
- Detail any different versions of digital content produced (such as high-quality or low-quality copies);
- Ensure that a complete description of how to use, reuse and/or reproduce the digital content—(preferably as a set of instructions)—is provided, as this will be critical for long-term preservation.

A good example of a DMP from the University of Cambridge is Dr Laurent Gatto's 'Data Management Plan for a Biotechnology and Biological Sciences Research Council (BBSRC) Tools and Resources Development Fund (TRDF) Grant' [25]. This DMP succinctly outlines information on the data, documentation, software, source code, and reproducible framework. Other examples of DMPs and guidance are available from the Digital Curation Centre [20], including the DMPonline tool [19].

## 12 Developing Documentation

As part of the transfer of custody of digital content outputs produced as part of a digital cultural heritage or digital humanities project, to a collecting institution or research repository, also supplying any associated materials is also essential. These associated materials will assist in comprehending the digital content's context and functionality. A project that produces an interactive online resource will need to ensure that the specifics of how different components connect is recorded. This may include documenting how any data flows through the system, including inputs and outputs (e.g. how data is entered/ingested, and how data is exported/made available). This information will be crucial for being able to preserve certain types of digital content in the long-term.

For digital content to be made available using future technologies on devices that are yet to be invented, a 'preservation action' may be necessary. This may take the form of a 'migration'[28] from one file format to another. For complex or interactive digital content, where dependencies between files exists, documenting these relationships between individual file components is critical. These relationships can be recorded as technical manifests, as well as other descriptive documentation (including within a DMP). If dependencies are unknown, a preservation action (such as a file format migration) may result in 'broken' digital content. For some complex interactive digital content, it may only be one part of the digital content that no longer works, however this may not be apparent if other aspects are still functioning. Given that preservation actions on digital content may be carried out in batches—without subject matter experts present, and without the ability to test every aspect of the migrated (or normalised) resultant version of the digital content—issues may not be discovered until it is encountered by a user, further down the track. For this reason, any complexities and dependencies should be thoroughly documented.

For example, as part of a transfer of custody for the DAAO eResearch tool, it would be necessary to provide the system diagram(s), information on system dependencies, metadata standards and other standards used, any data dictionaries implemented, controlled lists, metadata crosswalks, and the data model [16] (illustrated in Fig. 3).

This suite of documents form a key resource for managing the DAAO eResearch tool. These documents would be essential for ongoing maintenance, as well as for guiding any future preservation activities.

## 13 Borrowing from Related Industries

While digital stewardship of large-scale digital cultural heritage and digital humanities data is in its relative infancy, reinventing the wheel is unnecessary. Other disciplines have developed practical methods and methods for handling complex set-ups; borrowing and adapting these approaches to align with various digital preservation activities is recommended. One relevant sector for this purpose is the performing arts.

Mandatory documents produced as part of a performing arts production include (but are not limited to):

- Technical rider—a list of technical requirements for staging the performance;
- Stage plot—the layout of the performers, props, instruments, audiovisual and technical equipment etc.;

[28] In this context, migration—or normalisation—means to change a file from one file format to another (more standardised or commonly used) file format. These types of 'preservation action' migrations are often carried out in batches. Undertaking migrations of this kind can be intensive—-particularly for complex digital content—as ensuring no loss of any data or functionality is of primary concern.

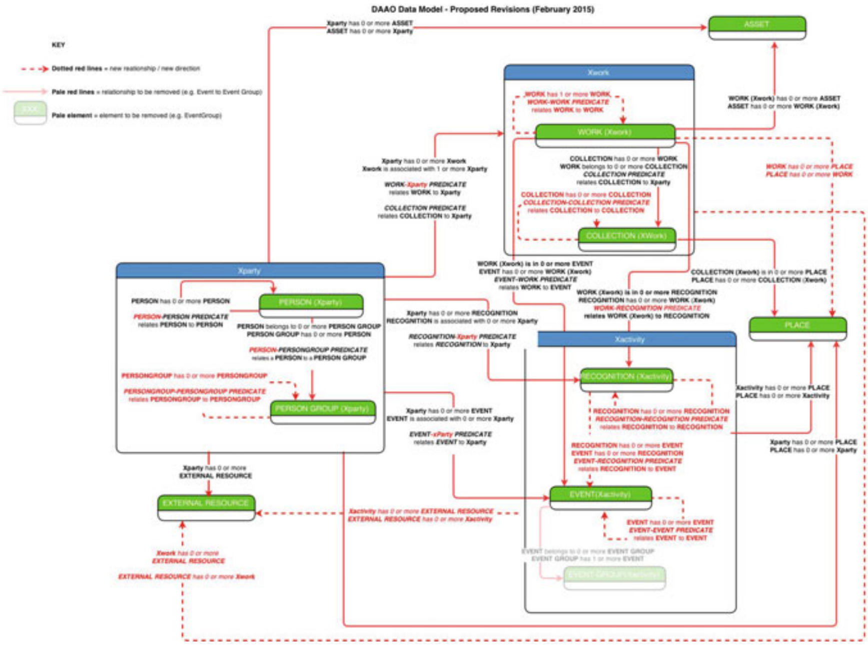

DAAO Data Model - Proposed Revisions (February 2015)
KEY
Dotted red lines = new relationship / new direction
Pale red lines = relationship to be removed (e.g. Event to Event Group)
XXX
Pale element = element to be removed (e.g. EventGroup)
Xparty has 0 or more ASSET
ASSET has 0 or more Xparty
ASSET
Xwork
WORK has 1 or more WORK
WORK-WORK PREDICATE
relates WORK to WORK
WORK (Xwork)
WORK (Xwork) has 0 or more ASSET
ASSET has 0 or more WORK (Xwork)
Xparty has 0 or more Xwork
Xwork is associated with 1 or more Xparty
WORK-Xparty PREDICATE
relates WORK to Xparty
COLLECTION PREDICATE
relates COLLECTION to Xparty
COLLECTION has 0 or more WORK
WORK belongs to 0 or more COLLECTION
COLLECTION PREDICATE
relates COLLECTION to Xparty
COLLECTION has 0 or more COLLECTION
COLLECTION-COLLECTION PREDICATE
relates COLLECTION to COLLECTION
COLLECTION (XWork)
WORK has 0 or more PLACE
PLACE has 0 or more WORK
Xparty
PERSON has 0 or more PERSON
PERSON-PERSON PREDICATE
relates PERSON to PERSON
PERSON (Xparty)
PERSON belongs to 0 or more PERSON GROUP
PERSON GROUP has 0 or more PERSON
PERSON-PERSONGROUP PREDICATE
relates a PERSON to a PERSON GROUP
PERSONGROUP has 0 or more PERSONGROUP
PERSONGROUP-PERSONGROUP PREDICATE
relates PERSONGROUP to PERSONGROUP
PERSON GROUP (Xparty)
WORK (Xwork) is in 0 or more EVENT
EVENT has 0 or more WORK (Xwork)
EVENT-WORK PREDICATE
relates WORK to EVENT
COLLECTION (Xwork) is in 0 or more PLACE
PLACE has 0 or more COLLECTION (Xwork)
WORK (Xwork) is in 0 or more RECOGNITION
RECOGNITION has 0 or more WORK (Xwork)
WORK-RECOGNITION PREDICATE
relates WORK (Xwork) to RECOGNITION
PLACE
Xparty has 0 or more RECOGNITION
RECOGNITION is associated with 0 or more Xparty
RECOGNITION-Xparty PREDICATE
relates RECOGNITION to Xparty
Xactivity
RECOGNITION (Xactivity)
Xactivity has 0 or more PLACE
PLACE has 0 or more Xactivity
RECOGNITION has 0 or more RECOGNITION
RECOGNITION-RECOGNITION PREDICATE
relates RECOGNITION to RECOGNITION
Xparty has 0 or more EVENT
EVENT is associated with 0 or more Xparty
EVENT-xParty PREDICATE
relates EVENT to Xparty
RECOGNITION has 0 or more EVENT
EVENT has 0 or more RECOGNITION
EVENT-RECOGNITION PREDICATE
relates RECOGNITION to EVENT
Xparty has 0 or more PLACE
PLACE has 0 or more Xparty
Xparty has 0 or more
EXTERNAL RESOURCE
EXTERNAL RESOURCE
EVENT(Xactivity)
Xactivity has 0 or more EXTERNAL RESOURCE
EXTERNAL RESOURCE has 0 or more Xactivity
EVENT has 0 or more EVENT
EVENT-EVENT PREDICATE
relates EVENT to EVENT
Xwork has 0 or more
EXTERNAL RESOURCE
EXTERNAL RESOURCE has 0 or more Xwork
EVENT belongs to 0 or more EVENT GROUP
EVENT GROUP has 1 or more EVENT
EVENT GROUP(Xactivity)

**Fig. 3** Design & Art Australia Online, 2015 data model

- Input list—an ordered list of the audio channels sent from microphones, instruments (via 'DI' boxes), computers and other electronic sound-producing devices etc., which are sent to a mixing desk.

These key documents for staging performing arts productions could be borrowed and then 'mapped' into the digital stewardship domain. Table 2 contains a proposed mapping between documents used in the performing arts and documentation required by the digital stewardship domain (incorporating digital curation, digital preservation and RDM).

Staging a professional performance without a tech rider would not occur in the performing arts. To support the digital stewardship of digital cultural heritage and digital humanities projects, when transferring custody of digital content outputs, these equivalent documents should also be considered as mandatory. In order to ensure this documentation is available during transfer of custody, it should be created early on in the project so as to record any complexities.

Some examples of technical riders that may be worthwhile referring to include a rider from the duo comprising American electronic music producer Grey Filastine and Indonesian 'neo-soul' vocalist Nova Ruth. Their 2017 live audiovisual performance, Drapetomania, incorporates traditional and electronic sounds and visuals [24]. Given the range of vocals and instrumentation (electronic and traditional) used in their performance—plus the digital video element—this is a comprehensive yet

**Table 2** Comparative proposed alignments between the performing arts and digital stewardship documentation

| Performing arts | Digital stewardship |
|---|---|
| Tech rider | Data Management Plan—laid out in list format, technical requirements etc. |
| Stage plot | System diagram, system dependencies diagram, data model etc. |
| Input list | Data flow diagram or list—information on inputs (entered/ingested) and outputs (exported/made available) |
| Other information held in the tech rider, and other production docs | All other associated documentation—standards used, metadata crosswalks, data dictionaries, controlled lists etc. |

condensed setup. Their tech rider clearly outlines the different components required to stage this performance, which may be a useful guide for digital cultural heritage or digital humanities projects that contain a wide range of elements. For complex digital content, the 2009 work, ZEE, by Austrian artist Kurt Hentschläger—who creates immersive audiovisual installations and performances—may prove a useful reference [27]. Hentschläger's tech rider clearly states which components are and are not supplied. Digital cultural heritage or digital humanities projects or outputs that contain multiple dependences will need to clearly identify each of these aspects. The tech rider for ZEE may be a useful documentation model.

Indigenous Australian contemporary dance company, Bangarra Dance Theatre's 2017 performance, Bennelong [52], incorporates a multitrack audio playback file (running out of QLab) with two data projectors 'blended' to create a giant projection on cloth, at the rear of the stage. Figure 4 illustrates the stage plot and input list, which form part of the tech rider for this production. In terms of a model for data flows—including inputs and outputs—a simple visual diagram, similar to the Bennelong tech rider, would be suitable for recording this type of information.

Given the need to capture precise technical information, as part of digital cultural heritage or digital humanities project outputs, or for reproducibility of research data, it is not that great of a leap between documenting technical information in the performing arts, to these domains. For projects that are nearing the end of their lifespans, this method of documentation allows for a meaningful way of detailing information, that will be necessary when undertaking preservation activities in future.

## 14 Conclusion

Taking a holistic digital stewardship approach allows for greater consideration of the available methods for capturing, managing and preserving digital content outputs,

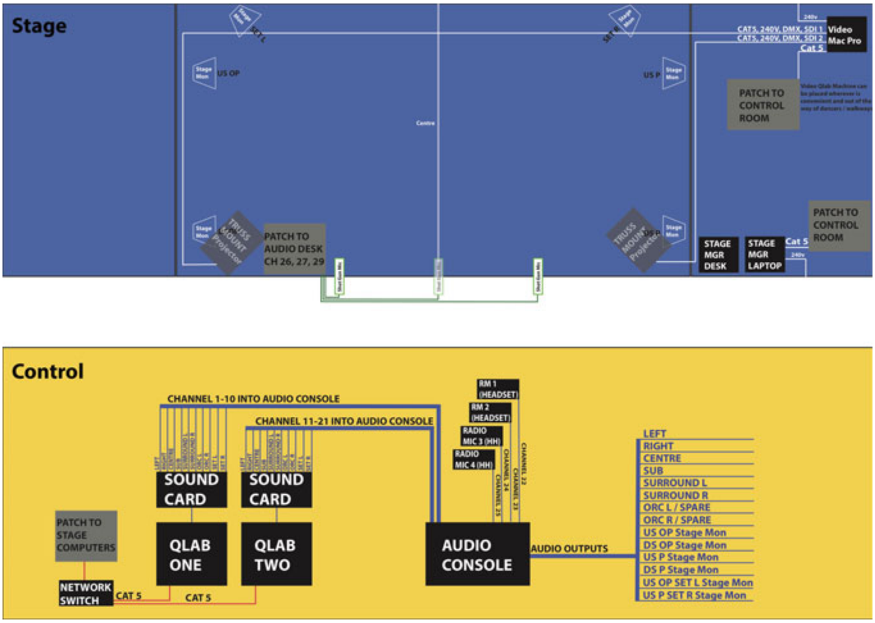


**Fig. 4** Stage plot and input list for Bangarra Dance Theatre's 2017 Bennelong production

created as part of digital cultural heritage or digital humanities projects. Developing a comprehensive awareness of the needs of each specific type of digital content, and producing documentation from the get-go, gives the digital content an improved chance of long-term sustainability. Plan the project's funeral early enough—ideally, as it is conceived—so that it allows for as many aspects of the digital content to 'live on'. Utilising several digital curation and digital preservation models—as well as implementing more than one capture approach in order to 'cover a few bases' (acknowledging there are pros and cons to each)—is likely to enable more options for the preservation of and access to digital content in the future. Considering the different ways in which future access to digital content may be offered will influence earlier preservation decisions. This will either provide future flexibility, or possible limitations downstream. Given other disciplines' approaches to capturing and documenting complex interactive digital content, and technical requirements and/or specifications, it would be pertinent for the field of digital stewardship (encompassing digital curation and digital preservation) to borrow and adapt selected approaches in order to record these technical complexities. There is no one perfect solution and the digital content outputs produced from each digital cultural heritage or digital humanities project, must be handled on a case-by-case basis, at least for the foreseeable future.

**Acknowledgements** The author would like to thank The Polonsky Foundation for their generous support in funding the Digital Preservation at Oxford and Cambridge (DPOC) project. Thanks to

Edith Halvarsson, Digital Preservation Specialist—Policy and Planning at Bodleian Libraries, and the other DPOC Fellows and Project Managers, at both the Universities of Cambridge and Oxford. Thanks also goes to Dr Andrew Thwaites, scientific director of the Kymata Atlas at the University of Cambridge, and Andy Lysle, Director of the Australian production company, Technical Events. Finally, thanks to Pascal Aeberhard.

## References

1. Brown, A.: Practical Digital Preservation: A how-to guide for organizations of any size. Facet Publishing, London (2013) 218
2. Bodleian Libraries, Oxford.: Digital Humanities Archives for Research Materials. (2014). Available at: http://blogs.bodleian.ox.ac.uk/dharma/2015/04/07/dharma-final-report/, accessed 27 March 2017
3. Cambridge Digital Humanities. Available at: https://www.cdh.cam.ac.uk/, accessed 28 May 2018
4. Cambridge University Libraries.: Digital Preservation Policy. Available at: https://doi.org/10.17863/CAM.32927, accessed 6 January 2019
5. Cambridge University Digital Library. Available at: https://cudl.lib.cam.ac.uk/, accessed 3 May 2018
6. Cambridge University Library.: Taylor-Schechter Cairo Genizah Collection. Available at: https://cudl.lib.cam.ac.uk/collections/genizah, accessed 27 March 2017
7. Cambridge University Library.: Classification. Available at: http://www.lib.cam.ac.uk/collections/classification, accessed 30 May 2018
8. Cambridge University Library.: Digital Humanities. Available at: http://www.lib.cam.ac.uk/research/digital-humanities, accessed 27 May 2018
9. Cambridge University Library.: Forman and Napier Casebooks. Available at: https://cudl.lib.cam.ac.uk/collections/casebooks/, accessed 28 May 2018
10. Cantara, L.: Long-term preservation of digital humanities. In: OCLC Systems & Services 22 (1). (2006) 38–42. Available at: https://www.researchgate.net/publication/220418724_Long-term_preservation_of_digital_humanities_scholarship, accessed 27 March 2017
11. The Casebooks Project. Available at: https://casebooks.lib.cam.ac.uk, accessed 2 May 2018
12. The CLOCKSS Archive. (2017). Available at: https://www.clockss.org/clockss/Home, accessed 2 May 2018
13. CLOCKSS.: Threats and Mitigations. Available at: https://documents.clockss.org/index.php?title=CLOCKSS:_Threats_and_Mitigations, accessed 16 June 2017
14. Data Hub of Australian Research on Marine and Aquatic Ecocultures. Available at: https://dharmae.research.uts.edu.au/, accessed 29 May 2018
15. Design & Art Australia Online. Available at: https://www.daao.org.au/, accessed 6 March 2018
16. Design & Art Australia Online Blog.: Schema and Data Model Refinements. (2015). Available at: http://blogs.unsw.edu.au/daao/blog/2015/03/schema-and-data-model-refinements/, accessed 30 May 2018
17. Dietrich, J.S.: E-Journals: Do-It-Yourself Publishing. In: Engineering and Science, Vol. 62 (4). (1999) 26–33. Available at: http://resolver.caltech.edu/CaltechES:62.4.EJournals, accessed 27 March 2017
18. Digital Curation Centre.: DCC Curation Lifecycle Model. Available at: http://www.dcc.ac.uk/resources/curation-lifecycle-model, accessed 29 March 2018
19. Digital Curation Centre.: DMPonline. Available at: https://dmponline.dcc.ac.uk/, accessed 30 May 2018
20. Digital Curation Centre.: Example DMPs and guidance. Available at: http://www.dcc.ac.uk/resources/data-management-plans/guidance-examples, accessed 30 May 2018

21. Digital Preservation at Oxford and Cambridge. (2016). Available at: http://www.dpoc.ac.uk/, accessed 27 March 2017
22. Digital Preservation Coalition.: Fixity and checksums. Available at: https://www.dpconline.org/handbook/technical-solutions-and-tools/fixity-and-checksums, accessed 25 May 2018
23. Dullaart, C, Sakrowski, R.: netart.database. Available at: http://net.artdatabase.org/, accessed 6 March 2018
24. Filastine & Nova.: Drapetomania Technical Rider. (2017). Available at: http://www.filastine.com/images/download/FilastineNova_Drapetomania_TECH_RIDER_15_Drapetomania.pdf, accessed 29 May 2018
25. Gatto, L.: Data Management Plan for a Biotechnology and Biological Sciences Research Council (BBSRC) Tools and Resources Development Fund (TRDF) Grant. (2017). Available at: https://doi.org/10.3897/rio.3.e11624, accessed 30 May 2018
26. GitHub. Available at: https://github.com/, accessed 2 May 2018
27. Hentschläger, K.: ZEE – Technical Rider 2009.1. (2009). Available at: http://www.kurthentschlager.com/portfolio/zee/techrider/Tech_Rider_ZEE_2009.pdf, accessed 28 May 2018
28. Hey, A J G, Trefethen, A E.: The Data Deluge: An e-Science Perspective. In: Berman, F, Fox, G C and Hey, A J G Grid Computing - Making the Global Infrastructure a Reality. (2003). Available at: https://eprints.soton.ac.uk/257648/, accessed 27 May 2018
29. Higgins, S.: Draft DCC Curation Lifecycle Model. (2008). Available at: https://doi.org/10.2218/ijdc.v2i2.30, accessed 2 May 2018
30. HTTrack Website Copier. Available at: https://www.httrack.com/, accessed 2 May 2018
31. International Internet Preservation Consortium.: OpenWayback. Available at: http://netpreserve.org/web-archiving/openwayback/, accessed 2 May 2018
32. Internet Archive.: Archive-It. Available at: https://archive-it.org/, accessed 7 May 2018
33. Internet Archive.: Heritrix. Available at: https://webarchive.jira.com/wiki/spaces/Heritrix, accessed 28 May 2018
34. Internet Archive.: Wayback Machine. Available at: https://archive.org/web/, accessed 17 March 2018
35. Johnston, L.: Digital Humanities and Digital Preservation (2013). In: The Signal. Available at: https://blogs.loc.gov/thesignal/2013/04/digital-humanities-and-digital-preservation/, accessed 27 March 2017
36. King's Digital Lab. Available at: https://www.kdl.kcl.ac.uk/, accessed 28 May 2018
37. King's Digital Lab.: What we do, Web and data hosting. Available at: https://www.kdl.kcl.ac.uk/what-we-do/web-and-data-hosting/, accessed 28 May 2018
38. Kenney A.R., McGovern N.Y.: Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems (2003). Available at: http://dpworkshop.org/dpm-eng/conclusion.html, accessed 27 March 2017
39. The Kymata Atlas.: (2017). Available at: https://kymata.org/, accessed 12 March 2018
40. Langley, S.: An approach to selecting case studies. (2017). Available at: http://www.dpoc.ac.uk/2017/05/26/selecting-case-studies/, accessed 27 March 2017
41. Langley, S.: Digital Preservation Should Be More Holistic: A Digital Stewardship Approach. In: Digital Preservation in Libraries: Preparing for a Sustainable Future. ALA Editions, Chicago (2018) 93–128. Available at: https://doi.org/10.17863/CAM.34317, accessed 6 January 2019
42. Langley, S.: Digital Streams Matrix (alpha release). (2018). Available at: https://doi.org/10.17863/CAM.26363, accessed 13 January 2019
43. Langley, S., Carter T., Davies, M., Gilmour I.: Managing multi-platform materials: selected case studies. In: Proceedings of the 19th International Symposium on Electronic Art, ISEA 2013 (2013). Available at: http://hdl.handle.net/2123/9706, accessed 27 March 2017
44. Lazorschak B.: Digital Preservation, Digital Curation, Digital Stewardship: What's in (Some) Names? In: The Signal. (2011). Available at: https://blogs.loc.gov/thesignal/2011/08/digital-preservation-digital-curation-digital-stewardship-what's-in-some-names/, accessed 27 March 2017

45. The Library of Congress.: Inaugural Class of National Digital Stewardship Residents Selected. (2013). Available at: https://www.loc.gov/item/prn-13-120/, accessed 14 May 2018
46. The Library of Congress.: National Digital Information Infrastructure and Preservation Program. Available at: http://www.digitalpreservation.gov/, accessed 12 March 2018
47. Lots of Copies Keep Stuff Safe. Available at: https://www.lockss.org/, accessed 29 April 2018
48. McCurry, J.: Digital Stewardship: The one with all the definitions. (2014). Available at: https://collation.folger.edu/2014/04/digital-stewardship-the-one-with-all-the-definitions, accessed 20 May 2018
49. National Digital Stewardship Alliance. Available at: https://www.ndsa.org, accessed 27 March 2018
50. Omeka. Available at: https://omeka.org/, accessed 4 May 2018
51. Ozmeka. Available at: https://github.com/ozmeka/ozmeka, accessed 4 May 2018
52. Page, S.: Bennelong.: Bangarra Dance Theatre. (2017). Available at: https://www.bangarra.com.au/whatson/productions/bennelong-2017, accessed 29 May 2018
53. Rhizome.: Oldweb.today. Available at: http://oldweb.today/, accessed 6 March 2018
54. Rhizome.: Webrecorder. Available at: https://webrecorder.io/, accessed 28 April 2018
55. Rosenthal, D S H et al.: Requirements for Digital Preservation Systems: A Bottom-Up Approach. (2005). Available at: http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html, accessed 24 May 2018
56. Software Independent Archiving of Relational Databases. Available at: https://github.com/DLMArchivalStandardsBoard/SIARD, accessed 30 March 2018
57. Swiss Federal Archives.: SIARD Suite. Available at: https://www.bar.admin.ch/bar/en/home/archiving/tools/siard-suite.html, accessed 28 April 2018
58. Thompson D.: Why digital preservation is or isn't business as usual. (2017). Available at: http://www.dpconline.org/blog/why-digital-preservation-is-or-isn-t-business-as-usual/, accessed 27 March 2017
59. United Nations Educational, Scientific and Cultural Organization.: Concept of Digital Heritage. (2017). Available at: http://www.unesco.org/new/en/communication-and-information/access-to-knowledge/preservation-of-documentary-heritage/digital-heritage/concept-of-digital-heritage/, accessed 2 April 2018
60. University of Cambridge.: Apollo Open Access Repository. Available at: https://www.repository.cam.ac.uk/, accessed 22 May 2018
61. UK Research and Innovation—Arts and Humanities Research Council.: Research Funding Guide. Available at: https://ahrc.ukri.org/funding/research/researchfundingguide/, accessed 6 January 2019
62. Video DownloadHelper. Available at: http://www.downloadhelper.net/, accessed 16 May 2018
63. Yang, H.: Total Cost of Ownership for Application Replatform by Open-source SW. (2016). Available at: https://doi.org/10.1016/j.procs.2016.07.170, accessed 16 April 2018