Walking in My Shoes: A Case Study from a Born-Digital Archive

Emmanuela Carbé
University of Pavia
emmanuela.carbe@unipv.it

Keywords: born-digital archive, digital curation, Italian contemporary literature, private papers archiving

Abstract

The vulnerability of bits and the obsolescence of media raise new challenges with respect to the preservation of cultural heritage produced in the last few decades. In 2009 a research team at the University of Pavia decided to develop the PAD (Pavia Archivi Digitali) project, aimed at the long-time preservation of digital papers from Italian writers and journalists and their accessibility to the research community. PAD is intended to be as flexible as possible in terms of types of material, numbers of authors and the dimensions of their archives: its main feature is an integrated quality control system that manages each single phase of a deposit almost in real time, allowing the ingestion, classification and validation of archives under strict and accurate supervision. The archival system is based on five areas: "staging", "deposit", "permanent", "work", and "info". The most difficult acquisition for PAD has been the archive of Francesco Pecoraro. It has been the best test case for procedures and workflow, for instance the ingestion of files from different media and of the materials published on his blog and on social networks. With the help of Pecoraro's archive, the PAD team designed software that facilitates cataloguing and managing digital archives.

Introduction

When the Italian journalist Beppe Severgnini submitted the proposal in 2009 to build a Born-Digital Archive of contemporary Italian authors, the University of Pavia did not have any idea how much of a challenge this would be. Shortly afterwards, Severgnini made available the developing PAD (Pavia Archivi Digitali) with more than 16,000 files from his own computer, and it soon became evident that the prototype project of archiving files of contemporary writers was on the point of becoming extremely complex and ambitious.

Nevertheless the University of Pavia had always been very attentive to new technologies and to the collaboration of different disciplines and so it seemed to be an ideal location to build a Born-Digital Archive – even more so because of its long-standing philological tradition. Back in 1969, Maria Corti came up with the ground-breaking idea of collecting manuscripts of twentieth-century Italian poets and novelists, and founded the Centre for Research in the Manuscript Tradition of Modern and Contemporary Authors, also known as Centro Manoscritti.

Consequently, thanks to the efforts of Professor Umberto Anselmi Tamburini (coordinator), Dr Primo Baldini (technical project and development) and Annalisa Doneda (responsible for interactions with the authors), a first working group was established in 2009. Currently PAD is chaired by Fabio Rugge, Chancellor of the University, and coordinated by Professor Paul Gabriele Weston. The Academic Board (http://pad.unipv.it/comitato) benefits from the work of many professors in various fields and areas. In addition, the staff of the University library system offer archival assistance and experience, and the attorney Luigi Ubertazzi and the legal office of the University of Pavia provide legal aid. The staff consist of two technical and scientific supervisors: Primo Baldini, who is in charge of the technical project; and Emmanuela Carbé, who supports development and testing of the software and liaises with authors. In the beginning, PAD could rely on the support of Fondazione Alma Mater Ticinensis of the University of Pavia, with the future objective of a profitable cooperation with the Centro Manoscritti.

PAD's mission is to collect and preserve born-digital materials provided by Italian authors, journalists and leading personalities in cultural fields: it consists of an archive of memories that contributes to the present Italian cultural landscape and which is easily accessible to the research community, complying with authors' privacy and copyright. After the original donation by Severgnini, five more authors have donated their archives to PAD: Silvia Avallone, Franco Buffoni, Gianrico Carofiglio, Paolo Di Paolo and Francesco Pecoraro. This has amounted to almost 80,000 files thus far. These authors vary greatly in age, education, and literary and journalistic approach and are highly diverse: this helps to build up a wide-ranging archive that is useful for the type of research that goes beyond just the literary sphere and provides samples of various methods of writing. The archives collected by the PAD project do not necessarily follow a schema and do not have any particular form: they sometimes contain files of different types, for example writing, graphics, media and documents generated using specific software.

Beyond paper

As we know, paper preservation is today only part of a wider problem. In February 2015, during the annual meeting of the American Association for the Advancement of Science, Vint Cerf, the Internet pioneer, addressed the vulnerability of memories that have been stored on digital platforms, which arise from the obsolescence of both hardware and software: what will twenty-first-century historians study? What strategies are in place to avoid the loss of the cultural heritage that has been created over the last few decades? Although the issue has been addressed previously (Kuny 1998), several problems remain unsolved to date: memory institutions face new challenges in securing the collective and personal memories of the last decades. The availability of large volumes of digital material raises questions about the role of digital curators in physical preservation and access to documents (Kirschenbaum, Ovenden, & Redwine 2010).

In 2010 Ricky Erway published a study which explained concisely and with extreme clarity a range of scenarios and issues about the long-term preservation of digital materials. Following that initial contribution, together with Barrera-Gomez in 2012 and 2013 she proposed certain fundamental steps required for the preservation of born-digital content extracted from physical media. They suggested the approach "walk before you can run". This is valuable advice for those who work in projects involving digital humanities, which rely on architectures based on scalability and interoperability. In the beginning, PAD looked like it would be a long journey and yet, despite the difficult experiences had with six authors and the improvement of all the procedures, every acquirement is characterized by new problems which are always different and unique.

The vulnerability of bits in fact has consequences in the field of literary archives: what kind of "manuscripts" have been produced by the writers of the last few decades? How is it possible to preserve today's "writers' desks" if nowadays everything is virtual, (only apparently) invisible, consisting of sequences of bits and binary code?

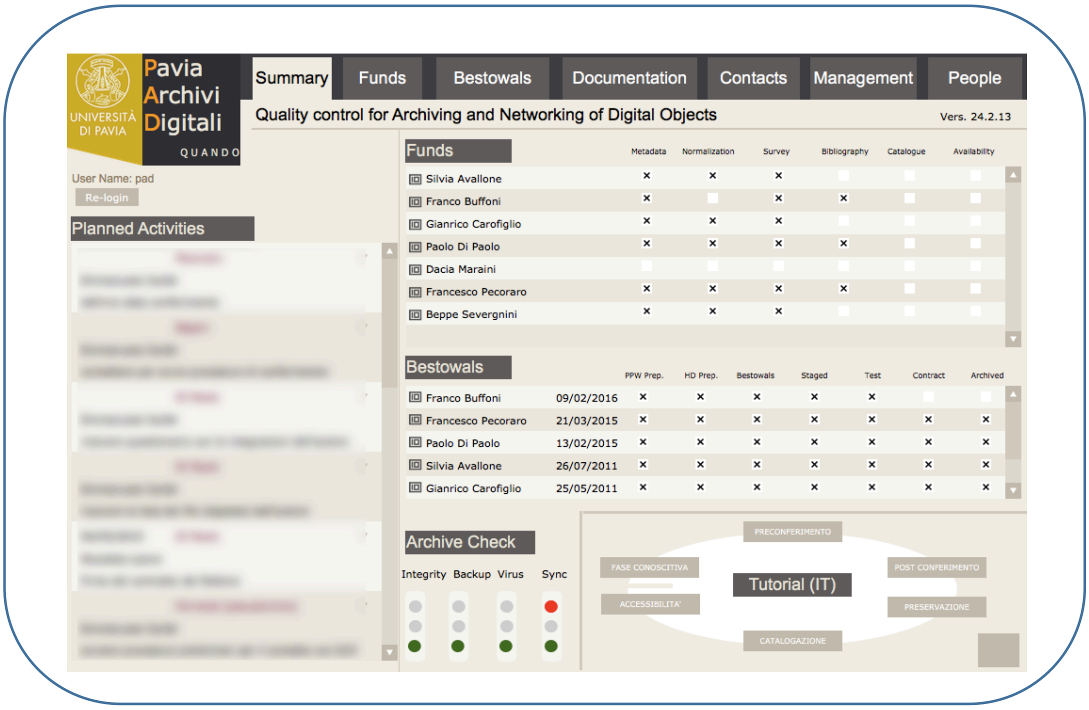
Only a few institutions have been working on projects for the preservation of born-digital writers' papers, including the Harry Ransom Center, which preserves collections such as that of Michael Joyce (Stollar Peters 2006). Another significant example is the collection of the Salman Rushdie digital archive, preserved by Emory University's Manuscript, Archives and Rare Book Library (Carroll, Farr, Hornsby, & Ranker 2011).

In the examples mentioned above, great effort has been put into ensuring the accessibility of the collections, for example by providing ways to emulate the original archive, and by the integration of paper and born-digital documents. Generally speaking, those few projects that emphasize literary archives focus on the works of a single author. The main goal of the PAD project is to implement a wider and more complex system dedicated to handling literary archives, with the added aims of comparing archives from several authors and incorporating a facility to parse texts with built-in textual analysis tools. PAD focuses on cataloguing the archives using adequate archival standards in order to ensure interoperability with traditional archives and perhaps with other born-digital archives that may, in the future, be more common than today. Along with some legal issues that are still to be settled, this is one of the major challenges in the project: given the amount of data that has to be handled, a fully manual cataloguing process would be unreasonable because the investment in terms of time and human resources would be too high. As a minimum, semi-automated and sustainable solutions are essential.

In these years of development of the project, PAD has amassed far more questions than answers regarding the management of writers' digital materials. We tried to explore different methods and perspectives, combining techniques that are typically used in other areas such as forensics, and applying them to the unique and specific features of a private literary archive with the intention of providing the DH community new questions to work on within a field that has received comparatively little substantive attention until now.

Methodologies and architecture

The aim of PAD is to be as flexible as possible in terms of types of content, number of authors and the size of their archives. So how should we want others to "walk in our shoes"? From the beginning, considerable effort has been put into the implementation of new technology and processes aimed at achieving better performance. Following the evaluation of established DAM solutions, the decision was taken in 2014 to develop an in-house software platform to seamlessly integrate with PAD's complex architecture. The software, entirely designed by Dr Primo Baldini, is called QUANDO (Quality control for Archiving and Networking of Digital Objects). The main tool in its development is FileMaker: the first version was initially a stand-alone and single-user program but, since 2015, it has been converted into a multi-user application that can be accessed within a private network (intranet). At the moment only staff can access the software using their personal credentials: users have different access levels, and can modify the data according to their roles within the project.



Quality control for Archiving and Networking of Digital Objects

Vers. 24.2.13

Funds	Metadata	Normalization	Survey	Bibliography	Catalogue	Availability
Silvia Avallone	☒	☒	☒	□	□	□
Franco Buffoni	☒	□	☒	☒	□	□
Gianrico Carofiglio	☒	☒	☒	□	□	□
Paolo Di Paolo	☒	☒	☒	☒	□	□
Dacia Maraini	□	□	□	□	□	□
Francesco Pecoraro	☒	☒	☒	☒	□	□
Beppe Severgnini	☒	☒	☒	□	□	□

Bestowals		PPW Prep.	HD Prep.	Bestowals	Staged	Test	Contract	Archived
Franco Buffoni	09/02/2016	☒	☒	☒	☒	☒	□	□
Francesco Pecoraro	21/03/2015	☒	☒	☒	☒	☒	☒	☒
Paolo Di Paolo	13/02/2015	☒	☒	☒	☒	☒	☒	☒
Silvia Avallone	26/07/2011	☒	☒	☒	☒	☒	☒	☒
Gianrico Carofiglio	25/05/2011	☒	☒	☒	☒	☒	☒	☒

Figure 1: Screen Summary of QUANDO

QUANDO manages all of the important aspects of the life cycle an archive and also acts as a repository for administrative documentation. It integrates information that has been entered manually with data that has been gathered automatically from other PAD software components (for checksumming, virus checking, metadata extraction, synchronization, etc.). It also assists in co-ordinating the efforts of all the various participants: the Academic Board, DAMS Administrator, Repository Administrator, staff, and legal consultants. The workflow can be coordinated through the software: from the first contact with the authors to the secure storage of data files (Weston, Carbé, & Baldini 2016).

The architecture of PAD has been designed in accordance with the OAIS Reference Model recommendations (CCSDS 2012; Lavoie 2014). The PAD archival system is based around six areas: "staging", "deposit", "work", "permanent", "info" and "database". The workflow procedure for every ingest requires the author to fill in an informational survey consisting of 15 main areas of questions. This is a primary step in the deposit process: we ask, for example, for information about the author's computer and devices, how the archive is organized, and how the authorship of the work and the technical tools relate to each other. This also helps provide more information about the relationship between author and computer, which is deeper than a purely technical relationship, and which can entail modifications to creative processes (Kirschenbaum, Farr, Kraus et. al. 2009): the process of acquisition of an archive can be intricate not only from a strictly technical point of view but also psychologically, since each author has a unique relation with the tools that she or he uses to write.

Upon initial ingest, materials are stored in the temporary area, where they are kept while waiting for the availability of an operator. In the deposit area, the integrity of the archive is checked, along with the possible presence of viruses. If any malware is found, the author is notified immediately and, if needed, assistance is offered. Viruses are usually quarantined within the PAD archive and they are only removed if a file could be irredeemably compromised, as described in the documentation: in such a situation, there are particular processes to be activated to try to recover the contents of a file. SHA-1 hashes are then generated. The PAD Print application generates a list of unique files that have been transferred, which is sent to the author for validation. In case of any second thoughts, the author can decide to remove a file or a set of files. Attached to the list of files is a summary indicating the total number of files transferred, the number of unique files and the size of the entire archive.

The work area is where metadata are extracted, documents are converted to formats that allow for better long-term accessibility and older computers may be emulated using virtualization technology. Finally, all of the data related to the deposit and collected by the QUANDO system are transferred to the "info" area. The database area has been created for the purposes of facilitating PAD workshops for the students of our university and to ensure accessibility for the research community. The permanent area is dedicated to preservation. An unencrypted copy of the archive is burned onto Gold Preservation archival standard DVDs and transferred to a bank vault. For every archive, two copies are stored in Pavia and another one in the University's facilities in Cremona, more than ninety kilometers away from the main site of the PAD project, thus following the principles of Distributed Digital Preservation (Skinner, Mevenkamp 2010).

A case study: Francesco Pecoraro's archive

The most difficult acquisition for PAD has been the archive of Francesco Pecoraro in 2015. It has been the best test case so far for our procedures and workflow, and has helped us to re-examine many aspects thereof, such as the ingestion of files from different media and of the posts published on the blog and on social networks.

The author, who is also an architect, was very popular for his blog “Tash-tego”, which was active from 2005 to 2011; he was subsequently quite active on Facebook until April 2015. He made his début with the short stories of *Dove credi di andare* (Pecoraro 2007), a collection of writings from his blog in *Questa e altre preistorie* (2009), the poems in *Primordio Vertebrale* (2011) and the novel *La vita in tempo di pace* (2013).

During our first meeting, Pecoraro noted that he first used a PC for writing in the 80s. He used to work with a workstation running Windows 7 at the time of the deposit and uses Dropbox for the majority of his writing. He also makes use of two external hard disks for storage, one of which also contains files that are not preserved on Dropbox: this hard disk is organized with directory names that informally describe where the files were previously located (for example: “White Thumb Drive”). The author backed up his work on a number of other occasions, especially (but not limited to) upon changing his workstation. His archive presented PAD with a "Russian doll" file structure and posed a number of problems in the first validation step.

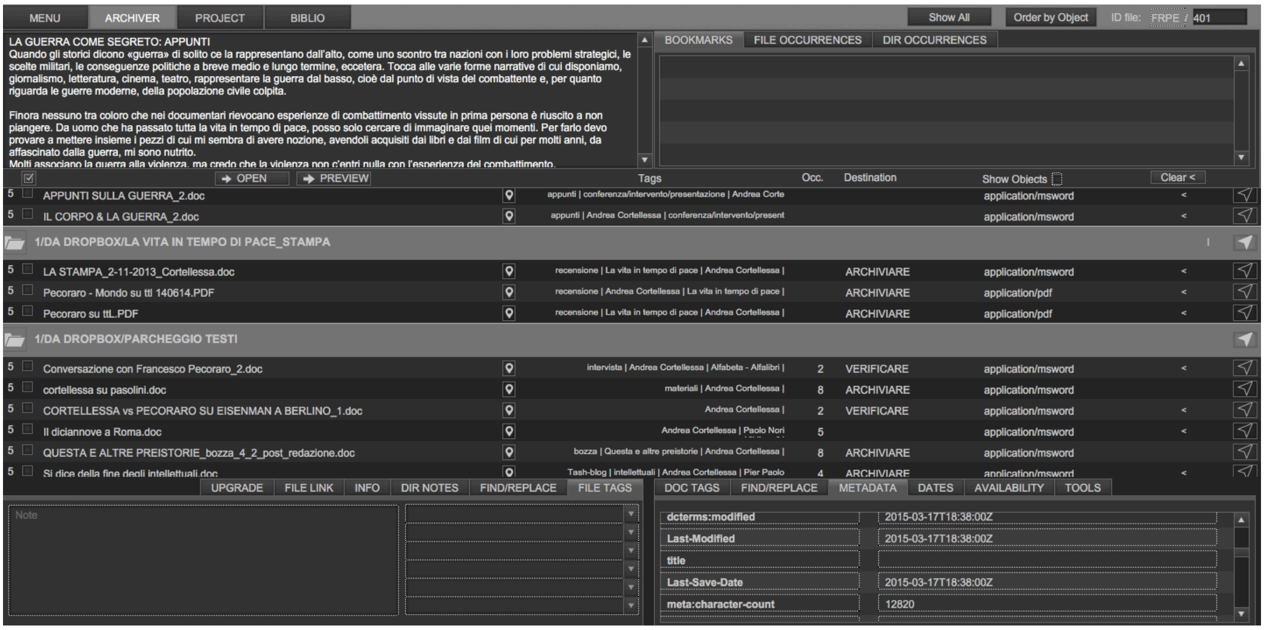
Pecoraro provided us with 35 floppy disks and the recommendation to convert any CAD files of his architectural work to either JPEG or PDF format. Of these 35 floppies, 10 could not be read any more and 5 of them contained a spanned ZIP file which could not be extracted even using an old version of WinZip95. The AutoCAD files have been converted to PDFs and, in the process, many viruses were found. Subsequently, we returned the materials to the author because PAD had refused to archive the obsolete media. He also gave us a DVD containing the work carried out by the editors of the book *Questa e altre preistorie*, including the final print version. We presented the author with a Kodak Preservation Gold DVD with all the files converted into open document formats.

After the first meeting we proceeded with the deposit process. The files were copied to a hard disk with hardware cryptography capabilities. We transferred files from Dropbox, the author's personal computer and an external hard drive. During the deposit process, the operator took notes and screenshots, taking especial note of the archival elements that would not be included in the deposit. We are always extremely careful in creating different folders on our external hard disk to represent the original source content and to consider them, from the archival point of view, only a reconstruction of the original situation.

Pecoraro gave us more than 43,000 files and specified that he wished to keep part of the private correspondence under embargo for 30 years. Obviously, in every ingestion, we always check in the temporary area to see whether an author has given us any highly personal files by mistake. One must face the problem, however, of how it is possible to check thousands of files to trace items that should be under embargo or highly personal files that need to be removed or kept unavailable.

From theory to practice: PADManager

This very difficult part of the project helped us focus on developing a system that could manage the file from the deposit to the permanent storage area, which would add metadata that could be helpful in the next steps of the process. Using Pecoraro's archive we designed PADManager, a piece of software that exploits the database functionality of PAD's architecture for cataloguing and managing archives. The future development of PADManager is projected to use techniques normally positioned in the field of artificial intelligence, such as machine learning and natural language processing. The ultimate aim is to allow the scientific community to access the archive and to provide tools for text mining and statistical analysis that could assist in the study of textual content.



LA GUERRA COME SEGRETO: APPUNTI
Quando gli storici dicono «guerra» di solito ce la rappresentano dall'alto, come uno scontro tra nazioni con i loro problemi strategici, le scelte militari, le conseguenze politiche a breve medio e lungo termine, eccetera. Tocca alle varie forme narrative di cui disponiamo, giornalismo, letteratura, cinema, teatro, rappresentare la guerra dal basso, cioè dal punto di vista del combattente e, per quanto riguarda le guerre moderne, della popolazione civile colpita.

Finora nessuno tra coloro che nei documentari rievocano esperienze di combattimento vissute in prima persona è riuscito a non piangere. Da uomo che ha passato tutta la vita in tempo di pace, posso solo cercare di immaginare quei momenti. Per farlo devo provare a mettere insieme i pezzi di cui mi sembra di avere nozione, avendoli acquisiti dai libri e dai film di cui per molti anni, da affascinato dalla guerra, mi sono nutrito.
Molti associano la guerra alla violenza, ma credo che la violenza non c'entri nulla con l'esperienza del combattimento.

☑			Tags	Occ.	Destination	Show Objects □	Clear <
5	□	APPUNTI SULLA GUERRA_2.doc	appunti | conferenza/intervento/presentazione | Andrea Corte			application/msword	<
5	□	IL CORPO & LA GUERRA_2.doc	appunti | Andrea Cortellessa | conferenza/intervento/present			application/msword	<
		1/DA DROPBOX/LA VITA IN TEMPO DI PACE_STAMPA					I
5	□	LA STAMPA_2-11-2013_Cortellessa.doc	recensione | La vita in tempo di pace | Andrea Cortellessa |		ARCHIVIARE	application/msword	<
5	□	Pecoraro - Mondo su ttl 140614.PDF	recensione | Andrea Cortellessa | La vita in tempo di pace |		ARCHIVIARE	application/pdf	<
5	□	Pecoraro su ttL.PDF	recensione | La vita in tempo di pace | Andrea Cortellessa |		ARCHIVIARE	application/pdf	<
		1/DA DROPBOX/PARCHEGGIO TESTI					
5	□	Conversazione con Francesco Pecoraro_2.doc	intervista | Andrea Cortellessa | Alfabeta - Alfalibri |	2	VERIFICARE	application/msword	<
5	□	cortellessa su pasolini.doc	materiali | Andrea Cortellessa |	8	ARCHIVIARE	application/msword	
5	□	CORTELLESSA vs PECORARO SU EISENMAN A BERLINO_1.doc	Andrea Cortellessa |	2	VERIFICARE	application/msword	<
5	□	Il diciannove a Roma.doc	Andrea Cortellessa | Paolo Nori	5		application/msword	<
5	□	QUESTA E ALTRE PREISTORIE_bozza_4_2_post_redazione.doc	bozza | Questa e altre preistorie | Andrea Cortellessa |	8	ARCHIVIARE	application/msword	
5	□	Si dice della fine degli intellettuali.doc	Tash-blog | intellettuali | Andrea Cortellessa | Pier Paolo	4	ARCHIVIARE	application/msword	<

dcterms:modified	2015-03-17T18:38:00Z
Last-Modified	2015-03-17T18:38:00Z
title	
Last-Save-Date	2015-03-17T18:38:00Z
meta:character-count	12820

Figure 2. PadManager: the *Archiver* Section

This test version is divided into three main sections: Archiver, Project and Biblio. Archiver, in particular, is the part that was developed while we were contending with the complexity of Pecoraro's archive. Each element can be checked through the software, where it is possible to select actions for all files and folders: checking the item with the author, identifying sensitive and undisclosable files, and locating files with technical problems. Every file can be seen in a simple preview or rendered to PDF format for ease of reading. There are functions to add temporary bookmarks to the archive, check technical metadata, and add tags either to a single document or to any instances of a document that is found replicated through the archive. Operators may also add chronological references, which are particularly useful when those that can be derived from the existing technical metadata appear to be incorrect with respect to the document contents. Bibliographic data collection is added in the Biblio section, which has been designed, for the time being, using a template that follows the Wikipedia guidelines in the hope of future publication as linked open data. Bibliographic data are needed for the archival description of funding sources (Project component of the platform), inspired by the FRBRoo model (Bekiari, Doerr, Le Bœuf, & Riva 2015).

The experience of ingesting a complex archive such as Pecoraro's showed us that there is still a long way to go in this work, not only in terms of the development and implementation of a data management application such as PADManager, but also regarding the improvement of the acquisition process, which does not only depend on expertise in Information Technology but also on the individual expertise of the operators. Hopefully there will be an opportunity to cooperate with other international institutions while we tread this path, with the common objective of improving best practice in the still relatively little known field of born-digital literary archive management.

REFERENCES

Barrera-Gomez, J., & Erway, R. (2013). Walk this Way: Detailed Steps for Transferring Born-Digital Content from Media You can Read In-house. Dublin, Ohio: OCLC Research.

Bekiari, C., Doerr, M., Le Bœuf, P., & Riva, P. (2015). *Definition of FRBRoo. A Conceptual Model for Bibliographic Information in Object-Oriented Formalism*. Den Haag: IFLA. https://www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo_v_2.4.pdf

Carroll, L., Farr, E, Hornsby, P., Ranker, B. (2011). A Comprehensive Approach to Born-Digital Archivers, *Archiviaria*, 71, 61-92.

Consultative Committee for Space Data Systems (2012). *Reference Model for an Open Archival Information System (OAIS)*. Washington DC: CCSDS Secretariat. https://public.ccsds.org/pubs/650x0m2.pdf

Erway, R. (2012). You've got to Walk Before You Can Run: First Steps for Managing Born-Digital Content Received on Physical Media. Dublin, Ohio: OCLC Research. http://www.oclc.org/research/publications/library/2012/2012-06.pdf

Kirschenbaum M. G., Farr, E. L, Kraus K. M. et al. (2009). Digital Materiality: preserving access to computer as complete environments, *iPRES 2009: The Sixth International Conference on Preservation of Digital Objects.* California Digital Library, 5-9 October, UC Office of the President, 105-112. https://escholarship.org/uc/item/7d3465vg

Kirschenbaum M. G., Ovenden, R., & Redwine G. (2010*). Digital Forensics and Born-Digital Content in Cultural Heritage Collections*, Washington DC: Council on Library and Information Resources.

Kuny, T. (1997). A Digital Dark Ages? Challenges in the Preservation of Electronic Information, *Workshop: Audiovisual and Multimedia joint with Preservation and Conservation, Information, Technology, Library Buldings and Equipment, and the PAC Core Programme*, 63rd IFLA Council and General Conference.

Lavoie, B. (2014). The Open Archival Information System (OAIS) Research Model: Introductory guide (2nd Edition). Dublin, Ohio: OCLC Research. http://dx.doi.org/10.7207/TWR14-02.

Pecoraro, F. (2007). *Dove credi di andare*. Milano: Mondadori.

Pecoraro, F. (2009). *Questa e altre preistorie*. Firenze: Le Lettere.

Pecoraro, F. (2011). *Primordio vertebrale*. Roma: Ponte Sisto.

Pecoraro, F. (2013*). La vita in tempo di pace*. Roma: Ponte Alle Grazie.

Sample, I. (2015). Google boss warns of “forgotten century” with email and photos at risk. *The Guardian*, 13th February. https://www.theguardian.com/technology/2015/feb/13/google-boss-warns-forgottencentury-email-photos-vint-cerf

Skinner, K., Mevenkamp, M. (2010). *DDP Architecture*, in Skinner, K., Schultz, M. *A Guide to Distributed Digital Preservation*. Atlanta: Educopia.

Stollar Peters C. (2006). When Not All Papers are Paper: A Case Study in Digital Archivy, *Journal of the Society of Georgia Archivists*, 24, 22-34.

Weston, P. G., Carbé E., & Baldini, P. (2016). Hold it All Together: a Case Study in Quality Control forn Born-Digital Archiving, *Qualitative and Quantitative Methods in Libraries (QQML*), 5, 695-710. http://www.qqml.net/papers/September_2016_Issue/5313QQML_Journal_2016_Westonetal_695-710.pdf