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

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

Keywords: Born-Digital Archive, Digital Curation, Italian Contemporary Literature, Private Papers Archiving

Introduction

When, in 2009, the Italian journalist Beppe Severgnini put forward the proposal to build a Born-Digital Archive of contemporary Italian authors, the University of Pavia did not have any idea that this would be a real challenge. Shortly afterwards, Severgnini provided the rising PAD – Pavia Archivi Digitali with more than 16,000 files from his own computer, and it soon became really clear that the local original project of archiving files of contemporary writers was about to become extremely complex and ambitious.

Nevertheless the University of Pavia had always been very attentive to new technologies and to the cohesion of different disciplines, and thus seemed to be an ideal location to build a Born-Digital Archive, also because of its long-standing philological tradition. Back in 1969 Maria Corti conceived the groundbreaking idea to collect manuscripts of twentieth-century Italian poets and novelists, and founded the Centre for Research in the Manuscript Tradition of Modern and Contemporary Authors.

Therefore in 2009, thanks to the efforts of Professor Umberto Anselmi Tamburini (coordinator), Dr Primo Baldini (technical project and development) and Annalisa Doneda (responsible for the interactions with the authors), a first working group was established. 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. Moreover, while the staff of the University Library System offers archival assistance and experience, the attorney Luigi Ubertazzi and the legal office of the University of Pavia provide legal aid. The staff consists of two technical and scientific supervisors: Primo Baldini, who is in charge of the technical project, and Emmanuela Carbé, who is supporting development and tests of the software and liases with authors. In its beginning PAD could rely on the support of Fondazione Alma Mater Ticinensis of the University of Pavia, with the future goal of a profitable
cooperation with the Fondo 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 can contribute to the definition of the Italian cultural landscape nowadays and that could be extensively accessible to the research community, complying with the author's privacy and copyrights. After Severgnini, five more authors have donated their archives to PAD: Silvia Avallone, Franco Buffoni, Gianrico Carofiglio, Paolo Di Paolo and Francesco Pecoraro. This has resulted in almost 80,000 files thus far. These authors, who greatly differ from one other in terms of age, education and literary and journalistic production, also represent highly diverse cases: this helps to build up a wide-ranging archive, useful for a type of research that goes beyond the mere literary sphere and provides samples of various methods of writing. The archives collected by the PAD project do not follow necessarily a schema and do not have a particular form: they sometimes contain files of different types, such as writings, graphics, media and particular application documents.

After the Paper

As we know, nowadays, paper preservation is only part of a bigger problem. In February 2015, during the annual meeting of the American Association for the Advancement of Science, the Internet Pioneer Vint Cerf addressed the issue of the vulnerability of memories that have been stored on digital support, which are subject to obsolescence of both hardware and software: what will twenty-first-century historians study? What are the strategies to avoid the loss of the cultural heritage that has been created over the last few decades? Although the issue had been addressed beforehand (Kuny 1998), several problems remain unsolved to date: memory institutions are involved in new challenges in securing collective and personal memories of the last decades. The availability of great amounts of digital material raises questions on 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 scenarios and issues about long-term preservation of digital materials. After that first work, she proposed in 2012 and 2013 with Barrera-Gomez certain fundamental steps for the preservation of born digital contents extracted from physical media. They suggested to "walk before running", and this is always a valuable advice for those who work in projects related to digital humanities, which rely on architectures based on scalability and interoperability. At the beginning, PAD – Pavia Archivi Digitali looked like a long walk and yet, despite the experience had with six authors and the improvement of all the procedures, every case is characterized by new problems which are always different and unique.

The vulnerability of bits has in fact 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 the new "writer's tables" 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 cases mentioned above, much efforts has been devoted to the inspectability of the collections, also by providing ways to emulate the original archive, as well as to the integration between paper and born-digital documents. Generally speaking the few projects that insist on literary archives focus on single cases. In the PAD project the main goal is to implement a wider and more complex system, dedicated to handling literary archives, also with the aim of comparing archives from several authors and to integrate the examination of materials with built-in text analysis tools. PAD focuses on the activity of cataloguing the archives using adequate archival standards in order to ensure the interoperability with traditional archives and possibly with other born-digital archives that will be more common than today in the future. Along with some of the 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 would be unreasonable, because the investment in terms of time and human resources would be too high. At least semi-automated and sustainable solutions are mandatory.

In these years of development of the project, PAD has collected far more questions than answers, regarding the management of writers' digital materials. We tried to explore different methods and points of view, combining techniques that are typically used in other areas, such as forensics, to the unique and specific features of a private literary archive, with the intention of providing the DH community new matters to work on in a field that has substantially received less attentions until now.

Methodologies and architecture

The aim of PAD is to be as flexible as possible in terms of types of material, number of authors and the dimensions of their archives. So how should we try to "walk in our shoes"? Since the beginning, a large effort has been devoted to the implementation of new technology and processes aiming at achieving better performance. After the evaluation of already established DAM solutions, in 2014 the decision has been taken to internally develop a software platform that could perfectly integrate with PAD's complex architecture. The software, entirely designed

by Dr Primo Baldini, is dubbed QUANDO – which stands for Quality control for Archiving and Networking of Digital Objects. The main tool used for its development is FileMaker: the first version used to be a stand-alone and single-user program, and since 2015 it has been converted to a multi-user application that can be accessed within a private network (Intranet). At the moment only the staff can access the software via their personal credentials: users have different access levels, and can modify the data according to their roles within the project.

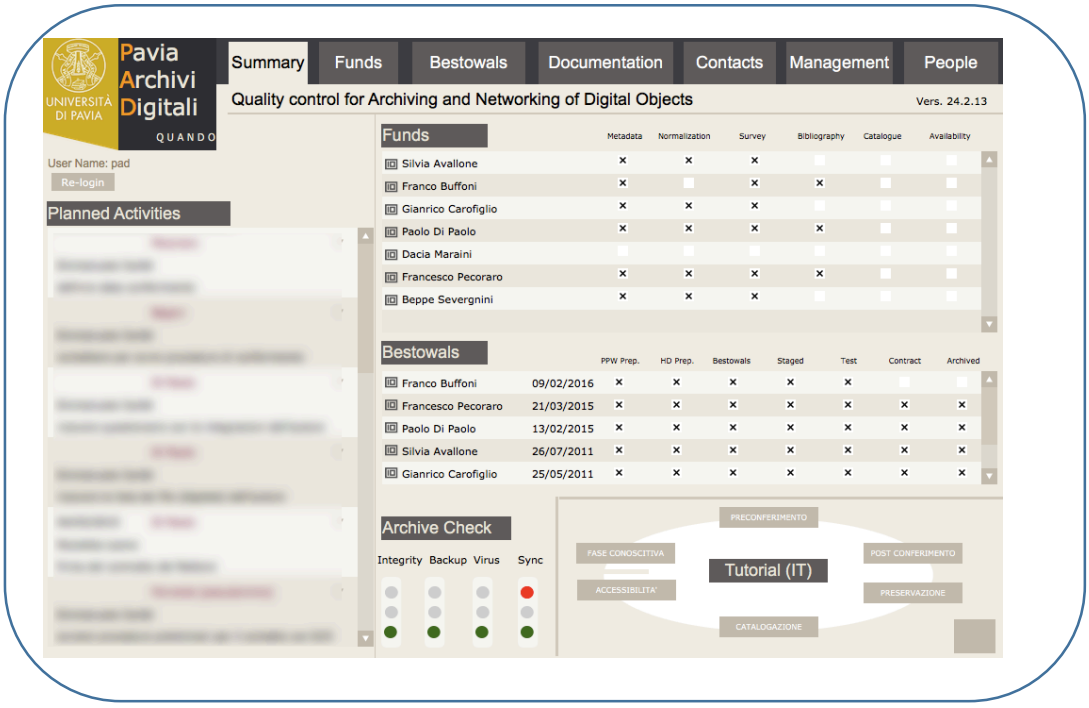

Figure 1. Screen Summary of QUANDO

QUANDO supervises all the important aspects of an archive's life, acting also as a repository for administrative documentation. It integrates information manually entered with data that has been gathered automatically from other PAD software (for checksumming, virus check, metadata extraction, synchronization, etc.). It helps in orchestrating all the involved personnel: 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 safe storage of data files (Weston, Carbé, & Baldini 2016).

The architecture of PAD has been designed following the OAIS Reference Model recommendations (CCSDS 2012; Lavoie 2014). The PAD archival system is based on six areas: staging, deposit, work, permanent, info and database. The workflow procedure for every ingest states to ask the author to fill in an informational survey consisting of 15 main areas of questions. This is a primary step for the bestowal itself, we ask for instance information about the author's computer and devices, how her or his archive is organized and how the work as a writer and the technological instruments relate to each other. This also helps to learn

something more about the relationship between an author and her or his computer, which is deeper than a purely technical one, and can entail changes in creative processes (Kirschenbaum, Farr, Kraus et. al. 2009): an archive's acquisition process can be delicate not only from a strictly technical point of view, but also on the psychological front, since each author has a unique relation with the tools that she or he uses to write.

Upon arrival, materials are stored into the temporary area, where they are preserved while waiting for the availability of an operator. In the deposit area the archive integrity is checked, as well as the possible presence of viruses. If any malware is found, the author is noticed immediately and, if needed, assistance is offered. Viruses are usually quarantined in the PAD archive; they are removed only if a file could be irredeemably compromised and this is reported in the documentation: in such a case there are particular processes to be activated to try to recover the file contents. Then SHA-1 hashes are 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 afterthoughts the author can decide to remove a file or a set of files. Attached to the list, a summary is sent indicating the total amount of transferred files, 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 a more durable availability and older computers are possibly emulated using virtualization technology. Finally, the whole documentation related to the bestowal and collected by the QUANDO system is transferred to the info area. The database area has been created to allow PAD workshops for the students of our university and for the accessibility to the research community. The permanent area is devoted to preservation. An unencrypted copy of the archive is burned on Gold Preservation quality DVDs and transferred to a vault located in a bank. 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 PAD main site, 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 that of Francesco Pecoraro's archive, which took place in 2015. It has been the best test case so far for our procedures and workflow, and helped us to reexamine many aspects, such as the ingestion of files from different media and of the materials published on the blog and on social networks.

The author, who is also an architect, was very popular for his blog "Tash-tego", active from 2005 to 2011; he was subsequently quite active on Facebook until April 2015. He debuted 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) which became a relevant literary case in 2014.

During our first meeting Pecoraro stated that he first used a PC for writing in the 80s. He used to work with a Windows 7 based desktop workstation at the time of the bestowal and uses Dropbox for the majority of his writings. He also makes use of two external hard disks to store materials, and one of them 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 more times, especially (but not limited to) upon workstation substitution. With his archive PAD faced in fact a chinese-box styled form of organization and several problems in the first validation step.

Pecoraro provided us 35 floppy disks with the recommendation to convert possible CAD files of his architectural work to either JPEG or PDF format. Among these 35 floppies, 10 could not be read anymore, and 5 of them contained a spanning ZIP file which could not be extracted even using an old WinZip95 installation. The AutoCAD files have been converted to PDFs and in the process many viruses were found. After that, we gave back the materials because PAD had decided to not archive the obsolescent media. He also gave us a DVD containing the work made by the editors on the book *Questa e altre preistorie*, and including the final version for the print. We donated to the author a Kodak Preservation Gold DVD with all the files converted in open document formats.

After the first meeting we proceeded for the bestowal. The files have been copied to a hardware cryptography capable hard disk. We transferred files from Dropbox, personal computer and an external hard drive. During the bestowal the operator took notes and screenshot about the context, especially noticing the archive parts that would not be included in the bestowal. We are always very careful in creating different folders in our external hard disk for the different original provenience and to consider them, from the archivistic point of view, only a reconstruction of the original situation.

Pecoraro gave us more than 43,000 files, and specified that he would keep part of the private correspondence undisclosed for 30 years. Obviously in every ingestion we always check, in the temporary area, if an author has given us some very private files by mistake. How is it possible, though, to check in thousands of files to trace undisclosed items or very private files that need to be removed or kept unavailable?

From the theory to the practice: PADManager

This very difficult case helped us focus on a system that could manage the file from the deposit to the permanent area, which would add metadata that could be helpful for the new steps of the process. With the help of Pecoraro's Archive we designed PADManager, a software that

exploits the database area of PAD's architecture and is devoted to cataloguing and managing archives. The future development of PADManager is supposed to use techniques normally positioned in the area 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 contribute to the examination of materials.

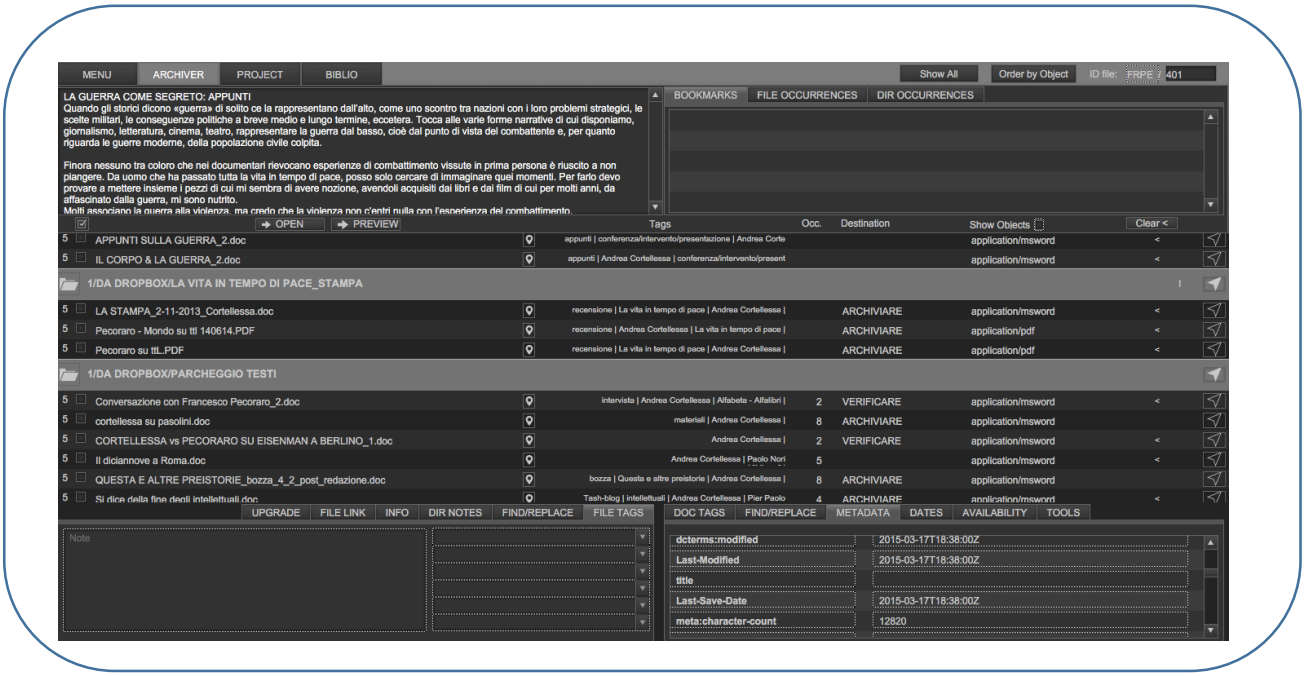

Figure 2. PadManager: the *Archiver* Section

This test version is divided into three main sections: Archiver, Project and Biblio. Archiver in particular is the part that has been developed while we were experiencing the complexity of Pecoraro's archive. Each element can be checked through the software, where it is possible to choose actions for all files and folder: checking with the author, determine sensitive and undisclosable files, find files with technical problems. Every file can be seen in a simple preview, or rendered to the PDF format for reading. It is possible to add temporary bookmarks in the archive, check technical metadata, add tags either to a single document or to all instances of a document that could be found replicated through the archive. It is also possible to add chronological references, which are particularly useful when the ones that can be derived from the existing technical metadata appear incorrect with respect to the document contents. Bibliographic data collection is added in the Biblio section, which has been designed, for the moment being, using a template that follows the guidelines of Wikipedia, in the hope of a future publication in the Linked Open Data. Bibliographic data are needed for the archivistic

description of the funds (Project section of the platform), inspired by the FRBRoo model (Bekiari, Doerr, Le Bœuf, & Riva 2015).

The experience with a complex archive such as Pecoraro's one showed us that there is still a long way to go, not only in terms of 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 Information Technology expertise but also on specific knowledge by the involved operators. Hopefully there will be the chance to cooperate with other international institutions while walking through this path, with the common goal to improve the best practices in the still not-well-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.