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

Emmanuela Carbé PAD –
Pavia Archivi Digitali
University of Pavia, Italy
emmanuela.carbe@unipv.it

Ricky Erway concisely explained scenarios and issues regarding the preservation of digital materials (Erway 2010), and with Barrera-Gomez proposed certain fundamental steps for the preservation of born digital contents extracted from physical media (Erway – Barrera-Gomez 2013). They suggest to "walk before run", a valuable advice for those who work in projects related to digital humanities, which relay on architectures based on scalability and interoperability.

The vulnerability of bits and the obsolescence of media also raise new challenges with respect to the preservation of cultural heritage produced in the last decades. The availability of great amounts of digital material of various kinds poses questions on the role of digital curators and memory institutions in physical preservation of digital material and accessibility to documents.

In 2009 a research team at the University of Pavia decided to develop the "PAD – Pavia Archivi Digitali" project, aiming at long-time preservation of digital papers from Italian writers and journalists, and their accessibility to the research community. Pavia seemed to be a good location to build a Born-Digital Archive, also because of the long-standing archivistic tradition of its "Centre for Research in the Manuscript Tradition of Modern and Contemporary Authors". PAD in its beginning consisted in a long walk and yet, despite the experience with six authors and the improvement of all the procedures, every case is characterized by new problems which are always different and unique.

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

The aim of PAD is to try to be as flexible as possible in terms of the amount of material types, authors and archive dimensions: its main feature is an integrated quality control system that manages each single phase of a bestowal almost in real time, allowing the ingestion, classification and validation of virtually every file type under a strict and accurate supervision. The locally developed Quality Control Software, dubbed QUANDO (Quality control for Archiving and Networking Digital Objects), is used to check all the important aspects of an archive's life, integrating information manually entered with data that has been gathered automatically using a PAD-developed application suite, which performs several actions on every single archive (checksumming, virus control, metadata extraction, synchronization, etc.).

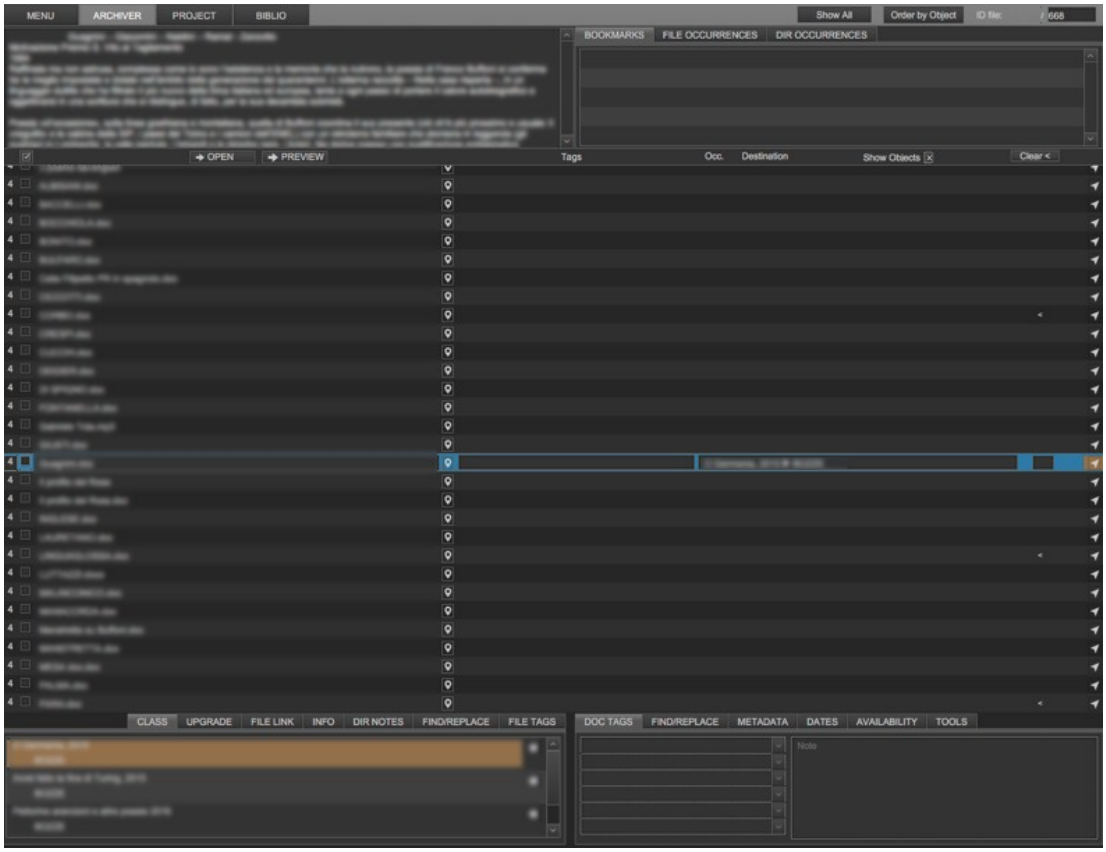
The most difficult acquisition for PAD has been that of Francesco Pecoraro's archive. It has been the best test case for our procedures and workflow, and helped us to reexamine many challenges, such as the ingestion of files from different media and of the materials published on the blog and on social networks. The author, who is also an architect, was very popular for his blog "Tash-tego", active from 2005 to 2011; thereafter he has been quite active on Facebook until April 2015. His most popular book is the novel *La vita in tempo di pace* (Pecoraro 2013), which became in 2014 a relevant literary case. In 2015 the author transferred more than 43.000 files to PAD, plus the materials coming from old floppy disks and a DVD. He gave the indication to keep some private correspondence undisclosed for 30 years. Before the bestowal, as stated in the workflow procedure for every ingestion, the author is asked to fill in an informational survey consisting of 15 main areas of questions. 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 keep materials. The author backed up his work several times, especially (but not limited to) upon workstation substitution. With his archive PAD faced in fact a chinese-box styled form of organization and a lot of problems in the first validation step, such as checking whether sensitive information was present in the archive and determining all files that should have been kept undisclosed.

The architecture for preservation and handling of data has been designed following the OAIS Reference Model indications. The archival system is based on five areas: staging, deposit, permanent, work, and info. Two copies of each archive, including documentation, metadata, and file format conversion, are kept in Pavia and another one is stored more than 90 kilometers away from PAD's main site.

Upon arrival, the PAD preservation plan prescribes that 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 possible viruses. In case any malware is found, the author is notified immediately and, in case of need, assistance is offered. SHA-1 hashes are generated. An 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. In the case of Pecoraro's archive many problems have been faced: how was it possible to check in more than 43.000 files, in terms of pointing out undisclosed items or very private files that have to be removed or kept unavailable, and then visually check complex cases such as this one? This very difficult experience helped us focus on an application that could manage the file from the deposit to the permanent area, which would add metadata that could be helpful for the fourth step of the process, in which the files are transferred to be catalogued. All Pecoraro's files are now being checked through this software, where it is possible to choose actions for all files and folder: such actions include checking with the author, determine sensitive and undiscloseable files, find files with technical problems and files to ask the author about.

In this way the author is be able to check and decide about all the archive files. At the same time it is possibile to add metadata, tag each files in two different ways (the single file or all the file duplicated), to add notes to files and directories, to review dates, to check the occurrences of files and directories, to preview each file and to directly open any file.



PAD software, developed using FileMaker

REFERENCES

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

Carroll L., Farr E., Hornsby P., Ranker B. (2011), A Comprehensive Approach to Born-Digital Archives, *Archiviara*, **72**, 61-92.

Erway, R. (2012) You've got to Walk Before You Can Run: First Steps for Managing Born-Digital Content Recevied on Physical Media. Dublin, Ohio: OCLC Research.

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

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