Long-term preservation of a web resource: PAD Web Archiving

Laura Pusterla[1], Primo Baldini[1], Paul Gabriele Weston[1]

[1]University of Pavia

Abstract: PAD-Pavia Digital Archives has recently developed a system aiming at the preservation of websites and social media pages. Through web scraping software, the author's website is made available locally, to guarantee offline browsing. Once the material is acquired, a disk image is generated, taking all the precautions to ensure disaster recovery. PAD also manages a series of collections of born-digital papers and if the relevant archive has already been described, it is possible to use it as an indexing tool for cataloguing the web files with an automatic cataloguing system.

Keywords: born digital archives, long-term preservation, web archiving, web preservation

1. Introduction

All of us have a daily relationship with the digital world and if we seek information, we know that we can almost always find it on the web. Although initially it might seem a trap of false news and biased data, for those who know where to look for on the internet it proves to be rich in great and important cultural sources. We entrust our information to the web because we are increasingly connected through our computers or our smartphones on an almost daily basis. In addition to social networks, people can choose to handle their own websites, while others increase or improve existing ones. Once dumped there, data and information are thought to be safe, stable and easily retrievable. Obviously, it's not that simple. Networked cultural heritage resources are, for several reasons, one of the most exposed to the risk of data loss. Network, storage or software malfunctioning, internal technical problems, and even hacker attacks, malware, hijacking accounts are just some of the dangers that could lead to loss of valuable information. In addition to these unintentional causes, internal reorganizations, that can modify the structure of the pages, or, on the contrary, poor maintenance of the site, enhance the risk of significant data oblivion. The recent unhappy fate of MySpace, among others, is there to prove it. Therefore, considering how many cultural testimonies of our age are present today only on the web, we can easily infer that the dispersion of these sites would cause a tremendous loss for our future generations. In this respect, the preservation of websites becomes fundamental when the latter are constituents of paper and digital libraries or archives.

2. PAD project

The PAD-Pavia Digital Archives project at the University of Pavia was launched in 2009 aiming at preventing born digital papers of contemporary authors from vanishing in the long term. The University hosts the Manuscripts Research

Centre, which has been preserving archives of Italian writers and journalists in paper format since 1969. It was the desire of the Centre to extend its expertise to digital native documents, based on the assumption that from now on all literary cultural production will be mainly based on the use of computer media. PAD preserves different types of materials, guarantees the long-term preservation of archival funds and will eventually provide access to scholars, in accordance with the provisions of the authors. So far, the project has been committed to the long-term preservation of digital archives hosted locally, that is on the personal computers and other devices owned by the authors. Lately, as a consequence of the growing trends in the use of networks and platforms, a system for the safeguard of websites and social media pages has been developed as an add-on. The intent of PAD is not to compete with other similar international projects, but as a small sustainable and high-quality project. The acquisition process must be initiated by the author or by the cultural institution that runs the site. In this way, we can interact directly with them to establish timing and methods for saving and assuring readability. All material of course remains property of the author that, at any time, may decide to remove it from the project.

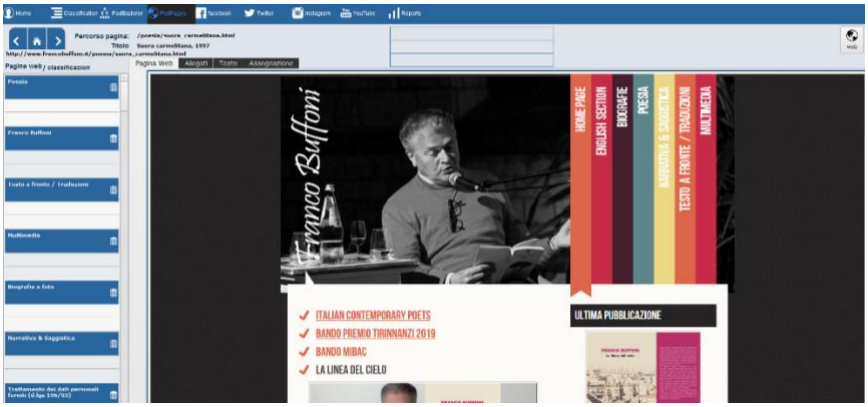
3. A case study: Franco Buffoni's website

In order to test the system, PAD used the website of the poet and anglicist Franco Buffoni (www.francobuffoni.it). The author had already given his computer files which had been catalogued using the PAD software.

In the first place, at the request of the site owner, PAD has picked up a copy. Since websites can be modified or updated very frequently, it may be preferable to agree on a policy of periodical pre-established downloads. In this way, it is possible to keep copies of the successive versions of the site, each of which will be available to scholars and readers. By using a software for web scraping, the author's Web site is transferred locally, so that it can be browsed offline. Thus, users can navigate freely the copy of the entire site. It can also occur that a previous version of the site has already been downloaded and that it is part of the archive which is bequeathed by the author. In this case the locally accessible site replaces the resource which is no longer available online.

The sites are sometimes quite complex structures, including on the one hand numerous references to other pages, either internal or external on the web, and on the other hand documents of various kinds, such as texts, images, video- or audio-recordings. For this reason, PAD not only stores every single page of the website, but also an image of those pages to which the various link lead, and the documents attached to the latter. By doing so, the system provides a better track of the path that the site creator wanted to achieve. For example, if a link to an external page is no longer working or the page no longer exists, part of what the author wanted to communicate goes missing. This operation is rather time consuming, given the amount of data that needs to be stored. Once all the material is downloaded and stored, a disk image is obtained, which is then saved in the PAD System

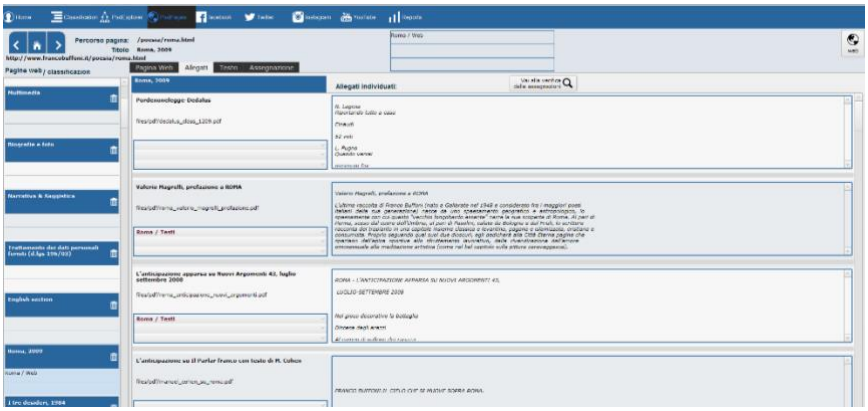
permanent storage area. All data are treated as digital native archives, and every precaution is taken to ensure disaster recovery. The protocol provides that all digital data are saved in triplicate: the first copy is hosted on the server at the University of Pavia; it is then mirrored on the server of the university branch at Cremona, whilst the last copy is burnt on a DVD-ROM. The same procedure is replicated for all saved versions. Another copy intended for access, is downloaded to PAD web server, which enables consultation from remote. In this way, given the author's consent, researchers and general users will be able to browse at leisure through the various versions of the site, using their personal credentials provided by PAD.



1 The navigation system of the site offline to PAD

When the software had to be tested, Franco Buffoni's web site was chosen as a testbed because of the variety in typologies and richness of material. Moreover, the description of its digital archive had already been completed making it possible to use this description as a reference point for cataloging the web component. It was chosen to focus on the textual resources, which are the majority in both the native digital archive and its web counterpart. By cataloging the archive bestowed by the author, we realized that often the author kept texts that were later transferred to the web, as either files attached to the page or links. In other cases, the texts of the web page have acted as inspirational sources, or they have been linked to other archival papers. By metadata extraction and then by considering the most important ones to identify the web resource in a timely manner, one can search for correlations between the digital native and the web. The process is not too different from the procedure that would normally be followed in treating material within PAD. At first, a site map is created, metadata are extracted and then reduced in extension by dropping those of little or no use,

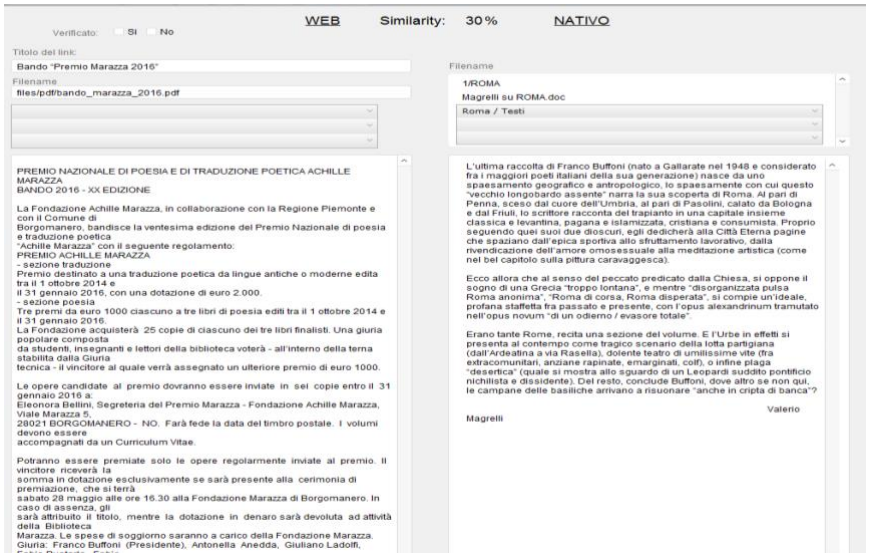
and finally the documents undergo an operation of normalization. At the end of this procedure, web documents can be analysed and described.



2 The display texts attached to page

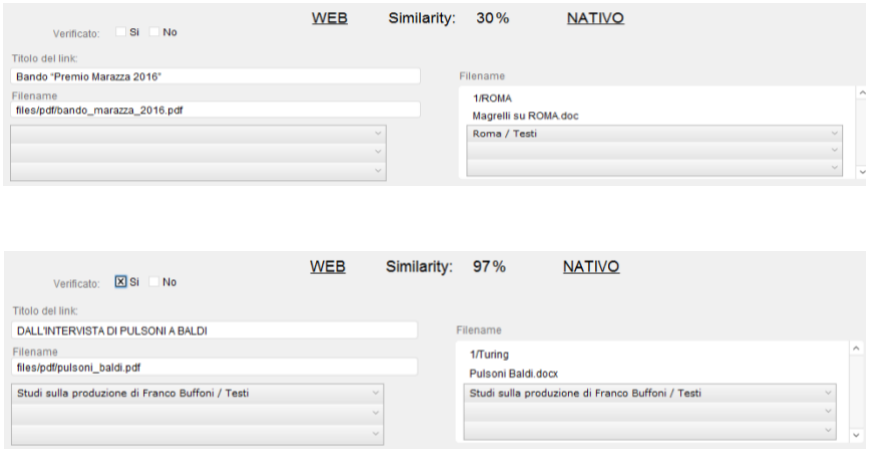
The cataloging system is automatic, which makes sense given the large amount of material available from the web that otherwise would be excessively long and time consuming to process, if it were a person to deal with it. Whereas a computer, after the command has been given, can process the material several hours non-stop, a human being would have to work for several days, in order to obtain a very similar result. The Web Analyzer software compares each web page, taking into consideration both the actual text of the page and possibly the files attached, with the existing documents in the archive. Then, it shows every web text faced with the text of the digital native files that most resembles. For each of them, a similarity index is established which indicates in percentage how much of the textual parts of the text on the web is identical to a document which is in the digital native archive. The index may provide a very low rate, under 50%, which would not be taken into account, or a higher one. If this were about 98%, the chances are that the compared documents have in fact the same contents. It must be considered that sometimes documents in the archive can be in formats quite different from those published or attached on the web (for example a Word file is usually converted to PDF in order to be placed on the site) and this brings as a rule a fall of the index rate. Obviously, the machine does not have the power of discernment typical of a human being. Therefore, the human factor is crucial in taking responsibilities along the various phases of the process. The possibility of a visual control on both texts in comparison has been anticipated. A software taking care of this specific function is in place to put the two texts next to one

other, so that the archivist can have an immediate view of both in order to dictate the correct way to catalogue them.



3 The comparison system

When a high index score is reported and therefore the computer infers that two identical files have been retrieved, it proposes the assignment to one of the inherent projects within the catalogue of the archive which was created manually by the archivist. Alternatively, with a low index score, the assignment box will remain blank and then the archivist will have to fill it in the manner it deems appropriate. Obviously, it is possible that either the text is produced directly on the network or that the original file was deleted, in which cases no matches are to be found.
To obtain these results PAD has implemented its *Levenshtein distance* algorithm software, adapting it to its needs.



4 In the first case a similarity index lower disallows any assignment. In the second we have the certainty that it is the same document, the system assigns it automatically.

4. Future perspectives

The PAD system for cataloging websites is constantly being modified and implemented, to provide new features and improve current ones. The choice to be configured as a small project, limited to sites of authors or cultural institutions, allows us to devote great care in fine-tuning the technical solutions according to the needs for archiving and cataloguing. The architecture of PAD Web Archiving has been designed for preservation, bearing in mind the need to make it easy for those who consult the catalogue and navigate to identify each component of the web resource. Therefore, when the site is saved, all the pages and attachments are systematically downloaded, allowing the user a navigation experience equal to the online one. In addition, the close collaboration with the authors ensures compliance with their decisions on the management of the website and allows to meet emerging needs. In order to meet social changes in the use of the Internet and its resources, PAD is currently experimenting a feature that allows long-term preservation and consultation of personal pages of social networks, such as Facebook, Twitter or Instagram.

References

Black, P. E. (2008), Levenshtein distance, *Dictionary of Algorithms and Data Structures [online]*. U.S. National Institute of Standards and Technology, retrieved 20/02/2019

Davis, R. C. (2016) Die Hard: The Impossible, Absolutely Essential Task of Saving the Web for Scholars. Eastern New York Academic and College Research Libraries Conference, Skidmore College, Saratoga Springs, NY.

Masanès, J. (2006). Web archiving: issues and methods. *Web Archiving*. Springer, Berlin, Heidelberg.

Weston, P. G., Carbé E. and Baldini P. (2017), Hold it All Together: a Case Study in Quality Control for Born-Digital Archiving, *Qualitative and Quantitative Methods in Libraries* 5.3, 695-710.

Weston, P. G., Carbé E. and Baldini P. (2017), If bits are not enough: preservation practices of the original contest for born digital literary archives. *Bibliothecae. it* 6.1, 154-177.