PaST Percorsi di Archeologia • 5

Coptic Literature in Context (4th-13th cent.)

Cultural Landscape, Literary Production, and Manuscript Archaeology

edited by
Paola Buzi

Edizioni Quasar

PaST

PAST – Percorsi, Strumenti e Temi di Archeologia

5

This volume, which is one of the scientific outcomes of the ERC Advanced project 'PAThs' – 'Tracking Papyrus and Parchment Paths: An Archaeological Atlas of Coptic Literature. Literary Texts in their Geographical Context: Production, Copying, Usage, Dissemination and Storage', has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 programme, grant no. 687567.

In copertina: *P. Mich.* 5421 e una veduta di Karanis

via Ajaccio 41-43, 00198 Roma - tel 0685358444
email: qn@edizioniquasar.it

eISBN 978-88-5491-058-4

Coptic Literature in Context (4th-13th cent.):

Cultural Landscape, Literary Production, and Manuscript Archaeology

Proceedings of the Third Conference of the ERC Project
"Tracking Papyrus and Parchment Paths: An Archaeological Atlas of Coptic Literature.
Literary Texts in their Geographical Context ('PAThs')".

edited by
Paola Buzi

Edizioni Quasar

Table of Contents

Defining Methodologies and Protocols for the Use and Reuse of Archaeological Legacy Data. The Case Study of the *Archaeological Atlas of the Coptic Literature*

*Julian Bogdani – Sapienza Università di Roma**

Abstract
This article deals with some methodological aspects related to the use of the so-called 'legacy data' and proposes some theoretical and practical points of reflection based on best practices that archaeologists can follow in their daily work with diverse data manipulation and creation. Moreover, the concept of 'legacy data' requires deeper consideration, to better define and contextualise its use within the scope of digital archaeological theory and practice. To offer a more concrete and solid framework to the general considerations, the case study of the *Archaeological Atlas of Coptic Literature* will be introduced, a research project that makes a consistent and systematic usage of data previously created by other projects, but pursuing rather different goals and having as a focus Coptic literary manuscripts. This project offers, in fact, some interesting illustrations of the different meanings that the term 'legacy data' may assume. Finally, a particular attention is paid to the transparent documentation of the research process and to the most efficient ways of publishing the documented dataset on the World Wide Web, in order to facilitate a scientifically consistent reuse of the data.

Keywords
'Legacy data', archaeology, manuscripts study, digital archaeology.

1. *Legacy data in archaeology: Theoretical, methodological and practical aspects*

The proper starting point for a discussion about 'legacy data' in the archaeological research is certainly the introduction that Penelope Allison wrote for the volume that she edited in 2008,[1] where the term is used for data that 'are not already digitised and geo-referenced, but must be prepared, and often manipulated, before they can be used in a digital environment'. Following this definition, and following the examples provided by the author, the Pausanias' Description of Greece and the non-digital records from the nineteenth and early twentieth century excavations at Troy can both be considered as 'legacy data'. Yet, in the ICT field, where the expression originated from, legacy data refers to data from obsolete information systems, in other words, data from information systems that are not in use anymore, although they might be in working condition. While it would be a stretch to consider the discursive form of a written book as an obsolete information system,[2] there are no doubts that excavations and documentation methodologies used in the past can be considered today as amply outdated and obsolete. Data contained in these systems need further elaboration and reconsideration before they can be compared to more recently collected ones or before they are being published in an integrated platform.[3] The volume edited by Penelope Allison deals mainly with aspects and issues concerning the digitising and the georeferencing of data preserved on paper supports[4] and does not consider issues related to the migration and reuse of already digital archives.

* The present article is one of the scientific outcomes of the ERC Advanced project 'PAThs' – 'Tracking Papyrus and Parchment Paths: An Archaeological Atlas of Coptic Literature. Literary Texts in their Geographical Context: Production, Copying, Usage, Dissemination and Storage', funded by the European Research Council, Horizon 2020 programme, project no. 687567 (PI: Paola Buzi, Sapienza Università di Roma), http://paths.uniroma1.it.

1 Allison 2008.

2 Strictly speaking, printed books and paper support is not to be considered as 'legacy' because it is still currently being used to document and publish archaeological data and research.

3 Jablonka 2004, 281-285.

4 On the important methodological issues involved in the process of digitising of 'analogue' cf. also Laužikas 2009, 247-259.

Fourteen years later, even the more technical aspect of recovering data from old or fallen into disuse systems is acquiring some importance and urgency for digital projects of various scales. More and more often the data that archaeologists need for their research are found in digital archives rather than on paper support, and the trend is certainly going to grow in the future since the documentation of the archaeological record is becoming gradually paperless.[5] The enthusiasm and bias in favour of digital approaches are paired with the fact that the use of technology has a cost, that might be negligible in the short-term but that might grow to prohibitive in the mid-to-long term. Digital data recording can be very fast and effective but there is no guarantee that after ten years we will be still able to read the files containing our data. The risk of becoming 'legacy' in a few years is more and more substantial. Yet, there seems to be no way back and digital information systems are today the only way to store and manage the data collected in the field and the libraries. On our side, we should aim at a greater awareness of these processes and try to apply to our everyday work the lessons we have learned by processing data from other sources. Moving information from legacy to cutting-edge information systems raises two different levels of problems, tightly connected. The first set of problems is merely technical and is about migrating the data, with possibly no loss. The second set of problems has a scientific base and lies on the foundations of the discipline because it regards the ever-changing informative potential of the archaeological (and broadly humanistic) record.

The technical aspect is mostly negligible, but it can grow to become a highly problematic issue. It is determined by external factors: a specific hardware or software may be discontinued by its producer, by consequently making not trivial the use, the recovery and the migration of data.[6] Depending on the degree of the obsolescence, the recovery operation might be strenuous and expensive and it is not surprising that the lack of funding causes in some cases the *sine die* deferral of the migration process, determining a *de facto* data loss. The access and the recovery depend therefore on the availability of technology and trained staff, in other words, funding, and the recovery process does not affect or change anyhow the content of the data.[7] Prevention is the most efficient strategy to avoid data loss. The careful planning of ordinary maintenance operations fit to delay the obsolescence of the information system and to facilitate and schedule regular migrations is something that must not be neglected. An informed choice of encoding formats, of software used and eventually of hardware infrastructure are mandatory requisites for efficient action.

The second perspective of legacy data treatment and reuse, brings into play the intimate nature of the archaeological record, a research theme that has produced a long-lasting discussion within the scholar community and has determined the definition of counterposed ways of thinking. The positivistic faith in the value *per se* of the archaeological record, much celebrated by the New Archaeology, has been opposed by the Post-processual (or contextual) Archaeology that doubts profoundly on the very possibility of totally objective observation of the archaeological record.[8] This point of view comes with the risk of excessively relativising the archaeological data and that of over-interpreting of the archaeological record, not considered 'simply evidence for the past, but are one of the media through which human beings constructed themselves and their communities in the past'.[9] On the other hand – as far as the automatised treatment of archaeological data is concerned – it is difficult not to agree with a 'neo-positivistic' (or neo-processual) point of view, which states that 'il dato archeologico, per sua natura, deve essere misurabile dal punto di vista quantitativo e qualitativo quando si intenda utilizzarlo in modelli di narrazione storica confutabili'.[10]

5 Ellis 2016.

6 A very classical example is that of the Domesday Project promoted by the BBC, that in the Eighties of the last century collected fragments of daily life of about one million people, to form a very rich archive containing texts and multimedia (images, video footage, audio clips, maps, virtual reality tours, etc.). It was published in 1986 and it was meant to be a snap-shot of the United Kingdom 900 years after the publication of the Domesday Book, the eleventh-century census of England ordered by William the Conqueror. In the Nineties, the advanced technology used to encode the information was outdated and the *ad hoc* developed memory devices were not supported anymore by the available hardware; the companies that had developed it had run out of business some years before. In 2011 some content was laboriously extracted from the original archive and published on a web site, that today has been archived (Cohen - Resenzweig 2006, 224-225).

7 Jeffrey 2012, 554.

8 Cf. Hodder 2005, 147-156. For a summary of the extensive bibliography on contextual archaeology cf. Barrett 2001, 141-164.

9 Moreland 2006, 139.

10 Valenti 2010, 8.

The 'agency' of the field archaeologist – that acts through his sensitivity, experience, and preparation – can be, to a certain degree, levelled out by applying stable standards and shared excavation and documentation methods. In any case, it is undeniable that the purposes and the special focus of the research project that is bringing to light the record, that is documenting, recording and analysing it, do condition the informative value of the data. This point is crucial and must be well kept in mind when the migration of data from legacy to up-to-date systems is performed. Not only technical aspects are involved in these processes, but also a sometimes profound transformation in the meaning of the content. By changing the research questions that have determined the creation of the archived records, we should expect that the information these data communicate might change. We should, therefore, be able to lead and govern this change and to acquire the maximum scientific profit from it.

A typical recovery operation of legacy data aimed at their reuse in more articulated and better-connected information systems allows not only the creation of faster and more efficient management, more fluid user experience and a more modern look and feel but also a new path to a better understanding of a specific domain. An old CAD drawing, for instance, extracted from the personal archive of the surveyor that compiled it, and integrated into a wider GIS platform offers a richer insight of the building as a whole. The same can be said for a database of potsherds or other archaeological finds migrated to a more complex database system that links the pottery fragments to other finds and the general archaeological context. New connections, new points of view and new research questions might bring to light new information from the available data.[11]

The migration and reuse of the archaeological datasets is never a neutral operation and this is a fact that archaeologists do know very well. The large use of previously collected data is, in fact, a fundamental part of every modern research project. Less common is the attitude of looking not only toward the past but also toward the future,[12] and consider how others will use the information we are digging today – both from the dirt and from the archives. In what manner future or present colleagues will be able to take apart and differently assemble our interpretative model, or more simply integrate our dataset in models of different geographic or cultural scale.[13]

The concern for the destiny of the digital research 'products' after the final report of funded projects has been for a long time not pressing, at least in the Italian context.[14] Even the funding institutions (mostly public), that one would expect to be seriously engaged in marking a clear trace of their activities into the society – not a *monumentum aere perennius*, but at least a feeble trace – have not determined mid-to-long term conservation policies of digital archives. It is the single researcher or project leader who decides on how and to what extent the archaeological record and the interpretative processes should be shared with the scientific community.[15] The underlying data – the 'objective' archaeological record – is fundamental, but particular importance holds also the interpretative framework used to collect and communicate these data. It is a sort of 'manual of use' a full description of technologies and methodologies used, that would permit a deeper comprehension and simpler reuse in the future. By documenting each step of the production process and the analysis of the archaeological record, as well as by documenting the technical data structure of our informative systems, we will be able to tremendously facilitate automated support renewal, the only effective way we have to extend the life of our (digital) archives.[16]

It is not easy to deal with vague notebooks or freehand sketches from the end of the 1800 or the first half of 1900, but archaeologists have well learned the lesson and the analysis and treatment of this kind of data is an important part of their methodology. It is not difficult to imagine how laborious could be in

11 While Contextual archaeology offers valuable theoretical ground for reflection, it seems too far to say that 'it no longer becomes possible to study an arbitrary defined aspect of the data on its own' (Moreland 2007, 83).

12 Cf. some important considerations matured in the context of the Archaeological Data Service in Jeffrey 2012, 553-570.

13 The transparency of every aspect of the interpretative model is a fundamental part of the ontology of the scientific research, that should be built on shared foundations following verifiable methodologies, Valenti 2000, 93-109; Bertoldi - Fronza - Valenti 2015, 233-243.

14 The trend is not only Italian, as the recent case of Trismegistos clearly demonstrates. Trismegistos has been for a long time a fully open access online database but funding issues have determined the setting up of a paid subscription plan: https://www.trismegistos.org/keeptrismegistosalive.php; for Trismegistos see below n. 23.

15 The sharing is not reserved only to peers, but a greater community can be interested and involved, as the Public Archaeology hopes. For an overview see Moshenska 2017b and particularly the introduction to this volume, Moshenska 2017a, 1-13.

16 Bogdani 2019, 120-121.

the future to deal with tens, or hundreds, or thousands of Gigabytes of undocumented and thus obscure archives, taking for granted that we would be able even to read them.

It is clear, hopefully, that the openness of the archives is an important step in their preservation and their ability to be easily reusable by others or by future us.[17] Moreover, a second fundamental step is their thorough and detailed documentation by using generic editorial platforms[18] and/or specific repositories and white papers. Metadata creation and documentation editing are time-consuming tasks, not rewarded from the academic system and not required from the funding system. Yet, this kind of grey literature is by far one of the best partners of raw data and together provide solid foundations for sound and durable research.[19]

2. *Case study: The* Archaeological Atlas of Coptic Literature

'PAThs: Tracking Papyrus and Parchment Paths. An Archaeological Atlas of Coptic Literature. Literary Texts in Their Original Context. Production, Copying, Usage, Dissemination and Storage' is an ERC funded project[20], whose goal is to provide an in-depth diachronic understanding and effective representation of the geography of Coptic literary production and in particular of the *corpus* of literary writings, almost exclusively of religious contents, produced in Egypt between the 3rd and the 13th century in the Coptic language. 'PAThs' combines together multiple disciplines, such as philology, codicology, palaeography, archaeology, archaeometry, and digital humanities in an effort to provide a detailed picture of the manuscript production in Coptic language of literary (i.e. not documentary) contents.[21] It provides a sufficiently complex and multifaceted case study to clearly illustrate some of the observations of the previous paragraphs.

Considering the long timespan (third-thirteenth centuries CE) and the huge geographic scope, it is clear that the project is heavily based on data previously edited and made available on paper and digital platforms. These data have undergone a very rigorous review and study process aimed at adapting them to answer new research questions, formulated by following the 'PAThs' research focus.

A detailed account would overflow the limits of this paper and its purpose. Only a few examples that better fit the general methodological frame already sketched above will be exposed in the following paragraphs. It is interesting to clearly and synthetically express from the beginning, some generic co-objectives of the 'PAThs' project, that might help a better understanding of the following examples.

- 'PAThs' aims to become a centralised publishing platform for data regarding the literary production in the Coptic language, representing a technological update of previous digital databases.
- The project is not a mere technological update, but formulates new research questions; for this reason, the data is being re-shaped to fit a better-connected scheme which allows a more detailed analysis of aspects previously not considered in detail.
- The *ad hoc* developed information system is designed to become a *multidisciplinary hub*, permitting the very detailed and deep analysis of specific branches, while maintaining a lucid connection net, fit to return an overall picture, by overstepping the high fragmentation of the research, a typical feature of our present.[22]

17 This is not an original or new idea: 'The lost cannot be recovered; but let us save what remains: not by vaults and locks which fence them from the public eye and use, in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident'. These are words part of a letter that Thomas Jefferson wrote to Ebenezer Hazard in 1791, the quotation is taken from Boyd - Lester 1974, 287-289.

18 Like the Journal of Open Archaeology Data https://openarchaeologydata.metajnl.com, for example, that publishes peer-reviewed data papers describing archaeology datasets with high reuse potential.

19 Dunning 2001.

20 Advanced Grant 2015, no. 687567. The project is directed by Paola Buzi and is based in Rome, at Sapienza University. More information is available at http://paths.uniroma1.it/.

21 Buzi 2017, 507-516; Buzi et al. 2017; Bogdani 2017, 59-69; Berno - Bogdani - Buzi 2018, 47-66; Bogdani 2018, 200-210. Since February 2019 a first version of the Atlas is available at https://atlas.paths-erc.eu.

22 While it is true that the over specialisation and the sometimes excessive fragmentation is a distinctive feature of the present-days research, this is a trend that plunges its roots deeply in the twentieth century: 'the knowledge of fragments, studied by turns, each for its own sake, will never produce the knowledge of the whole; it will not even produce that of fragments themselves', Bloch 1953, 155.

- Finally, 'PAThs' aims to become an automated endpoint for the distribution of structured, well-documented and scientifically reliable data, encouraging collaboration and promoting the full transparency of the entire scientific process.

2.1 Coptic manuscripts from CMCL to 'PAThs': Data transformation and migration

As already mentioned, 'PAThs' is by no means the first online database dedicated to the recording and study of the manuscript tradition in the Coptic language. On one hand, larger projects focused on manuscript tradition also consider books written in Coptic.[23] On the other hand, the pioneering work of Tito Orlandi and his *Corpus dei Manoscritti Copti Letterari* (CMCL) has determined a fundamental advance both in the field of Digital Humanities and on that of Coptic studies.[24]

From the technical point of view, the CMCL is a flat-file database, loosely following the relational model. The data are stored in text files and the management and query logic is implemented by using Unix scripts: no programming languages other than the Unix Shell is used. The same philosophy is followed for the web publication of the data, that are directly sent via the Unix command to the Common Gateway Interface (CGI). This extremely simple configuration grants to the CMCL an incomparable speed, efficiency and durability in time.[25] One of the drawbacks of this architecture is the lack of validation logic during the data-entry phase, but this has never been an issue since Tito Orlandi is the only authorised user who can write on the database.

With regards to the content, the database assigns and maintains a series of unique identifiers for bibliological units, textual units, author units, and narrative units and defines policy for their interconnection.[26] This classification is the result of many decades of meticulous work by Tito Orlandi and his scientific collaborators, who have carefully analysed a tremendous amount of manuscript fragments in the attempt to recover the original codicological units (i.e. books) and to try to narrate their history and the history of the literary works they contain.

The migration of the data contained in the CMCL database to the new 'PAThs' database was an important prerequisite for the construction of 'PAThs' *Archaeological Atlas of Coptic Literature*. Even if the CMCL database is still available on the Internet (and hopefully will be maintained from many years to come) the migration process took the shape of the recovery of legacy data. A relational-like, flat-file database system with no support for indexing and no data-validation policy had to be programmatically transformed into a fully relational, SQL based database system and several issues had to be addressed.[27] The most challenging problems, on the other hand, were related to the particular needs of 'PAThs' that required a custom data structure, different from one designed for CMCL. Three examples can be enlightening and representative of the tangled research and technical problems that the migration and reuse of legacy datasets in newly created information systems might bring into the light.

2.1.1 Author units and textual units' connection

Each author unit (i.e. author) filed in the CMCL makes one or more connections to textual units (i.e. literary works), following a one-to-many pattern, which is an over-simplification of the state of art of our

23 It is the case of Trismegistos, https://www.trismegistos.org/, Depauw - Gheldof 2014, 40-52

24 The CMCL database is available at http://www.cmcl.it through a paid subscription plan but some sections are made freely available. For detailed information on the database cf. Orlandi 2003.

25 The only external dependency of this platform is the Unix operating system and the Unix-shell that contains the few text manipulation utilities based on regular expressions used by the CMCL (basically *sed*, *awk*, *grep* and similar). I own this fundamental information on CMCL to personal communications by Tito Orlandi who generously shared the founding philosophy of this information system that he is still maintaining.

26 Orlandi 2008, 7-12.

27 Among others, the issue of text encoding had to be faced. T. Orlandi had developed his own system for encoding Coptic script using plain ASCII characters; the system had been inspired by the Beta Code invented by David W.Packard in the late 1970s and used to encode Greek texts by the *Thesaurus Linguae Grecae* http://stephanus.tlg.uci.edu/encoding.php. A software library was specifically developed to transform Orlandi's encoding to standard Unicode and was later released with an open-source license. The library, named cmcl2unicode, written in Vanilla JavaScript, is hosted on GitHub (code: https://github.com/paths-erc/cmcl2unicode, demonstration: https://paths-erc.eu/cmcl2unicode/ and indexed in Zenodo https://zenodo.org/badge/latestdoi/76262299.

knowledge on Coptic literature and manuscript tradition. This uncomplicated connection was intended by Orlandi as a mere symbolic representation of the reality and was determined by the technical difficulty of realising and maintaining many-to-many labelled relationship within a flat-file database. The technical issue was overcame by the use of a modern relational database managing system, which made possible and simplified a more articulated link between works and authors, by using a many-to-many qualified (labelled) relationship.[28] This case makes clear to what extent a technical issue might condition and limit the understanding of the data, and it also firmly stresses the importance of the metadata in the consideration of a structured dataset: Tito Orlandi (and other scholars using his CMCL) was well aware of the complexity of the authorship-related issues in the Coptic literature, albeit this was not fully represented in his data structure.

2.1.2. List of identified Coptic manuscripts: Study-first vs. evidence-first approach

Another significative difference between CMCL and 'PAThs' is the naming policy of the codicological units, i.e. ancient books. Even though the concept behind the 'PAThs' entity labelled *Manuscripts* fully coincides with the already mentioned Codicological Units of the CMCL, 'PAThs' has assumed a completely different approach for their naming. T. Orlandi provided with a *siglum* each new unit that he or his collaborators were able to identify and reconstruct with certainty. Furthermore, the *siglum* was indicative of the bibliological unit (i.e. 'groups of codices having formed a library at some time in the antiquity')[29] where the manuscript originally was contained and was based on a hexavigesimal system.[30] There exists also a bibliological *siglum* for manuscripts of unknown provenance (CMCL, e.g. CMCL.AA). Manuscript fragments that have not been yet attributed to a re-established ancient book are excluded until a deeper study can better define them. This approach could be described as 'study-first' and perfectly reflects the history of the CMCL, the lifetime work of a passionate scholar and his collaborators. 'PAThs', on the other hand, is a fixed-term (5 years) project aiming at creating a long-lasting and sustainable infrastructure. It was thus imperative to create from the beginning a clear map of the material evidence of Coptic manuscript tradition, in order or be able to easily design the work of the next decades. This is the reason behind the decision to provide each fragment excluded from the CMCL classification with a codicological unit identifier, following an 'evidence-first' approach. While at a first glance it would have been rational to expand the numeration system set up by CMCL, it was decided to create a new primary key, called CLM,[31] which is susceptible to future changes as far as the isolated fragments are concerned.[32] This breaking change, despite potentially confusing,[33] proved to be

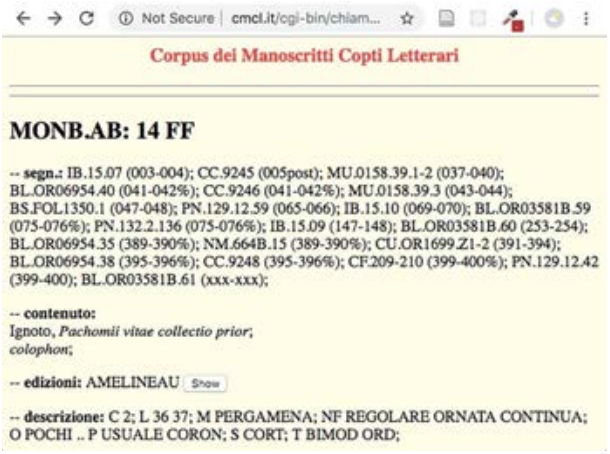

Fig. 1. A screenshot from CMCL showing the available information on codicological unit MONB.AB.

28 A work may be reported in a manuscript to have been written by an author (who is labelled as a 'Stated author') even if we know that it has been actually written by a different author (labelled as 'Creator'). For a detailed discussion of what is meant by 'Stated author' and 'Creator', cf. BERNO - BOGDANI - BUZI 2018, 51. The new information system keeps detailed track of this complexity.

29 http://www.cmcl.it/cgi-bin/chiamata.cgi?ms@codici=Show.

30 For example, the codex named MONB.HF was recognised to be part of the library of the White Monastery at Atripe (MONB = Monastero Bianco). The unique identifier is expressed by two digits of hexavigesimal system, using the letters of the English alphabet. This intuitive system maintains a very compact and fixed-size length and allows a very high number of unique identifiers (24^{24} = 6.156196E36).

31 CLM stands for Coptic Literary Manuscript.

32 A realistic scenario is the merging of many CLMs, recognised after careful study to belong to the same codicological unit. The suppressed CLM numbers are not reused and a map of the changes through time is maintained and published. Manuscripts bearing a CMCL identifier are more stable (theoretically immutable) while the others still need further analysis.

33 The scientific community, understandably, is not very comfortable with the introduction of new naming systems and identifiers, especially when the old ones are widely known and used in publications.

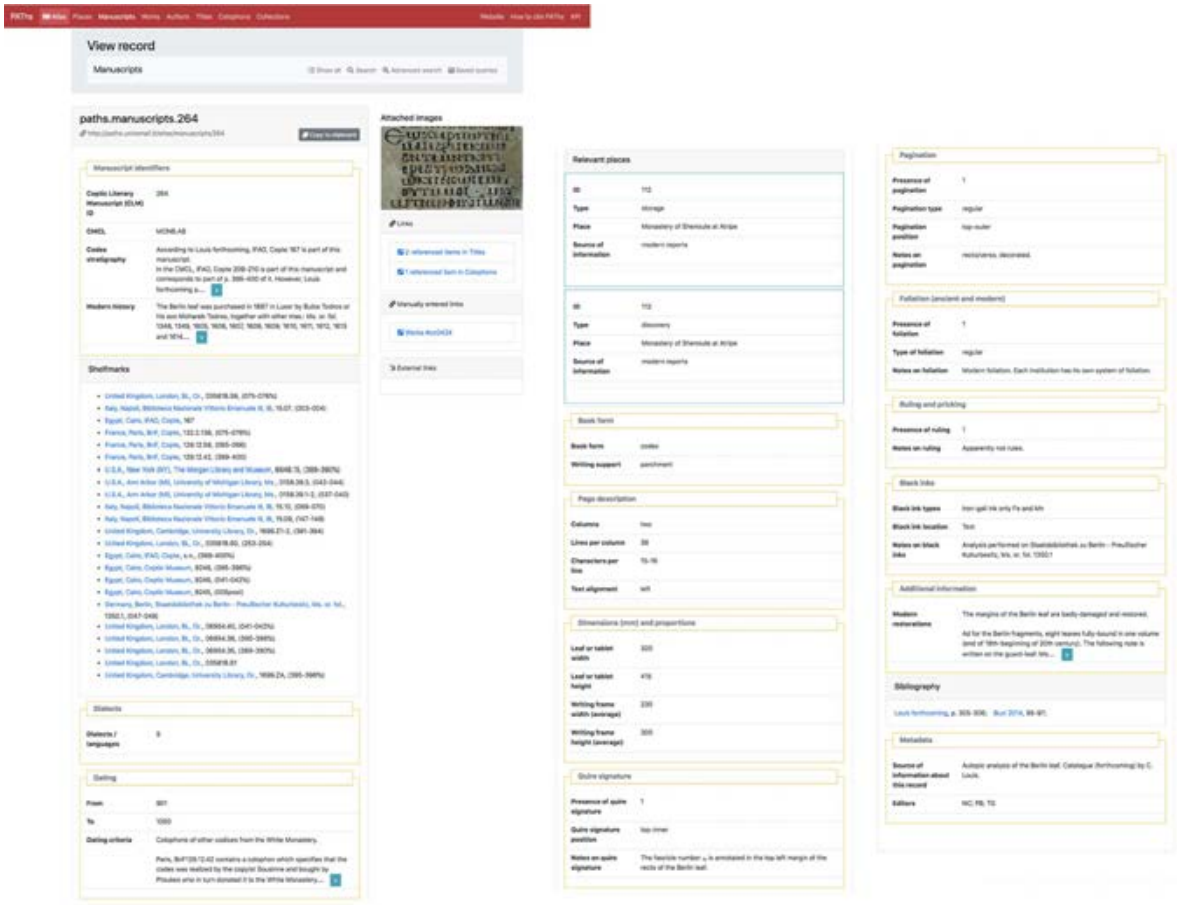

Fig. 2. A screenshot from 'PAThs' *Atlas* (https://atlas.paths-erc.eu/manuscripts/264) showing the available information on codicological unit CLM264 (=MONB.AB).

strictly necessary due to the partially different goals of the CMCL and 'PAThs' project. The corresponding 'CMCL' identifier is provided with each CLM entry, to offer to scholars a complete and updated map of the available resources.

2.1.3. Extension of the descriptive protocol

The main aim of the CMCL database was to univocally identify books, libraries, authors, and works, by providing a neat and bare map of the several entities. Typically, a manuscript is described by its *siglum*, the list of the extant fragments (usually scattered in several modern collections), the list of the works it contains, the list of the bibliographic records containing its edition and possibly a concise description. Any other information is to be found, if available, in the reference bibliography. 'PAThs' has considerably extended this part by developing a new protocol aimed at providing a thorough description of the codicological aspects of the ancient books.[34] This addition represents an important original contribution by 'PAThs' and broadens our perspective on the material aspect of the manuscripts, a concern shared in general terms by Tito Orlandi's work, but that had never found a place in the CMCL.

These few but representative methodological considerations are an integral part of the inner structure of 'PAThs' and as such have been fully described in the documentation that is published and kept up to date in a separate dedicated platform,[35] whose development is going in parallel with the *Atlas*.

34 For an overview of this protocol cf. the manual of use published at https://docs.paths-erc.eu/handbook/manuscripts.

35 https://docs.paths-erc.eu.

2.2 Methodologies and protocols for a geodatabase of Christian Egypt

Just like the codicological description, the archaeological contextualisation and the geographical representation are part of PATHs' contribute to Coptic studies.[36] By now, an almost final version of the *Archaeological Atlas of Coptic Literature* has been made freely available[37] and submitted to the scientific community for review and feedback. The rich web application offers full access to the entire dataset developed by 'PATHs' through very intuitive and user-friendly interfaces, that give access to simple and detailed search functionalities. Also to some pre-compiled queries (called saved queries) have been prepared, to suggest to general public some interesting starting points of consultation and to provide relevant examples to the research questions that can be asked to the dataset. The application includes also a geographical interface – the proper *Atlas* – able to represent on a map the distribution of archaeological sites (*Places*) considered to be important for the comprehension of Late Antique Egypt.[38] The web GIS makes wide use of available cartographic resources, both representing the current situation (base cartography and satellite imagery) and the ancient one.[39]

Once more, this was the starting point of a still on-going process, aimed at producing a rich and open geographic database, with a specific Egyptian focus by trying to acquire, digitise, georeferenced and publish online[40] maps and cartographic coverage of different dates and editions. Particularly, the georeferencing process, that inevitably deforms the original map, do highlight important matters on how to relate to the original and how to deal with inaccuracy.

> What one should keep in mind is that georeferencing does not necessarily improve a historical map or make it more accurate. In the course of changing the original map to make it amenable to digital integration, georeferencing changes lines and shapes, the distance between objects, the map's aesthetics, and its value as a cultural artefact. One gains knowledge of the original while processing it for inclusion in GIS, but one also loses something if the original map is not represented for comparison with its actual size, proportions, and qualities. Ideally, researchers should include both the warped map and the scanned image of the original map in a GIS project or publication.[41]

2.2.1 Georeferencing legacy archaeological and architectonic graphical documentation

The issues introduced above and related to the general methodological framework are currently being addressed and surely will require a deeper analysis and a publication on their own.[42] A GIS-related ex-

36 For an introduction to the archaeological review of Late Antique Egypt, still underway, cf. the articles by Angelo Colonna and Ilaria Rossetti in this volume.

37 The *Atlas* was published at https://atlas.paths-erc.eu on the occasion of the Third 'PATHs' International Conference held at the Sapienza University of Rome between 25 and 27 February 2019. The web application is still under development and so are the underlying data. Yet a great effort has been made to provide from the very first steps the main functionality. The *Atlas* is developed with exclusive use of client-side technologies (it is based on the React JavaScript framework, https://reactjs.org; the project code is hosted on GitHub: https://github.com/paths-erc/atlas and is indexed in Zenodo https://zenodo.org/badge/latestdoi/140484435. The dataset is not packaged and shipped with the application, but it is retrieved in real-time from the central web database via a REST API, following a workflow already successfully tested in other similar applications, cf. Bogdani 2016, 236-245.

38 Bogdani 2017; Bogdani 2018, 204-206. For an updated description of the type of archaeological sites (*Places*) included in the dataset, cf. the introduction available at https://atlas.paths-erc.eu/places.

39 In the recent years the archaeologic research is increasingly experimenting web technologies to deliver and distribute GIS data to a wider public, encouraging the reuse and stimulating the collaboration. For the most used technologies and protocols employed to publish geographical data, cf. Bogdani 2019, 97-102; Previtali - Valente 2019, 17-27. At present (2020), 'PATHs' is using generic basemaps distributed by Google through the Google Maps Platform https://cloud.google.com/maps-platform/, present-days topographic coverage distributed by Open Street Maps https://www.openstreetmap.org/, a map of the Roman Empire distributed by the Digital Atlas of Roman Empire at the Centre for Digital Humanities, University of Gothenburg, Sweden https://dh.gu.se/dare/, https://web.archive.org/web/20191022004307/http://commons.pelagios.org/2012/09/a-digital-map-of-the-roman-empire/ and a physical map of the Ancient Mediterranean area distributed by the Ancient World Mapping Center http://awmc.unc.edu/.

40 If and when copyright permits the redistribution and the creation of derivate work.

41 Rumsey - William 2002, 6.

42 'PATHs' is dealing with a great number of different cartographic coverages, designed from the late eighteenth century to the present-day. A provisional index is available at https://docs.paths-erc.eu/data/. The maps are mostly available in Public Domain license, downloaded from the Perry-Castañeda Library Map Collection of the University of Texas at Austin, https://legacy.lib.utexas.edu/maps/. Other images have been donated by the friends of the Polish Centre of Mediterranean Archaeology of the University of Warsaw (PCMA). The coverage of the Egyptian territory is highly uneven, but hopefully, the pursuit of the work will

ample of legacy data transformation and reuse might be of some interest here, regards the efforts that are being spent to create a comprehensive and georeferenced archive of drawings and maps of churches, basilicas and other buildings related to the Christian religion of Late Antique and Medieval Egypt. This geo-archive is being integrated into the *Archaeological Atlas of Coptic Literature*. The main sources for this archive are the many drawings, maps, and sketches that have appeared in the archaeological literature, but also topographical surveys – available on paper, as raster images or as vector graphics – that partner missions operating on the field are wishing to share.

Fig. 3. Comparison of the coordinates of the same, clearly recognisable ground control point: the base of the southern pylon of the temenos of Taposiris Magna (paths.place.338). The upper image is imported from Bing and the lower one from Google satellite images. Coordinates are expressed in the same projection system, WGS 84/ UTM zone 36N (EPSG: 32636). The linear difference can be calculated to more than 5 meters.

Before all, the important work of Peter Grossmann must be pointed out. His indefatigable on-field efforts have documented and shared with the scholar community an impressive number of Christian buildings, many of them now partially destroyed or buried under the sands. The book,[43] and the many articles he has written are often an exceptional and unique source for many lost archaeological contexts.

The workflow is canonical. Each drawing or map of a religious building is firstly digitised and a full reference to the bibliographic item where it was published is filed along with the digitised image.[44] Then, it is georeferenced using Desktop GIS software and non-deforming algorithms.[45] Legacy architectonic graphical documentation (typically, drawings of single buildings or, in most fortunate cases, maps of groups of buildings) do not usually contain any indisputable element to clearly define their position on the Earth's surface. These documents do not report normally grids of coordinates or topographical stations of known position. In the Fifties of the last century, GPS did not exist and cartographic coverage was not always at disposal, and surely not at a scale that we would judge acceptable today. In most cases, the drawings do report the north indication (it is unclear if it is the magnetic or the cartographic north) and a scale bar. In most cases, the building is reported to be located in a site of known location, although the relative position is mostly lacking.

fill most of the gaps. Special attention is being drawn on the 'masterpiece of cartographic compilation and early nineteenth-century fieldwork' to cite the title of a book dedicated to them (Godlewska 1988): 42 maps that Pierre Jacotin compiled under the guidance of the Napoleonic expedition in Egypt (Jacotin 1824).

43 Grossmann 2002.

44 PAThs' Zotero group, https://www.zotero.org/groups/2189557/erc-paths/items, is being used as a unified repository for all bibliographic references.

45 The GIS platform is developed using the QGIS open-source software http://qgis.org. During the georeferencing process, the *Linear* and the *Helmert* transformation types have been used, since they do not distort the original raster as the various *Polynomial* and *Thin Plate Spline* do. The *Freehand raster georeferencer plugin* for QGIS (code available at https://github.com/gvellut/FreehandRasterGeoreferencer) proved to be a valuable tool, inasmuch it provides graphical tools to perform the basic operations while the official *Georeferencer* tool requires pre-extracted coordinates.

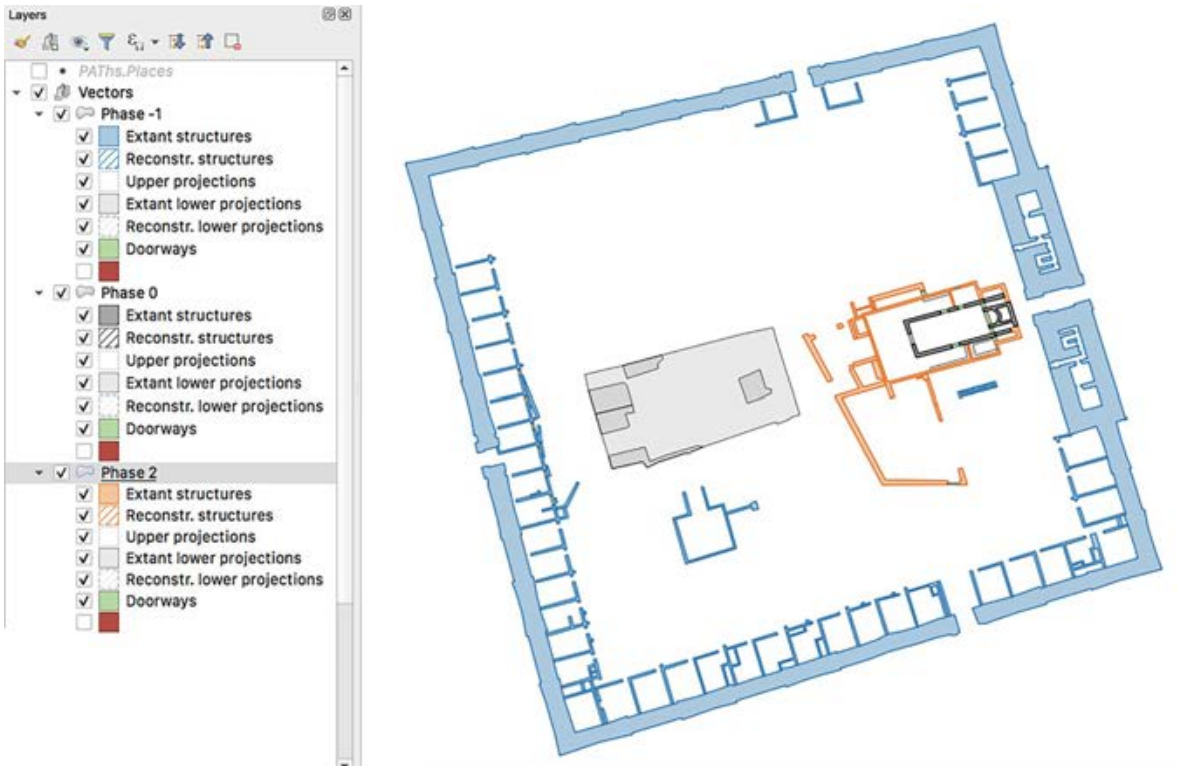

Fig. 4. Example of the use of the SVP, the temple and church of Taposiris Magna in a GIS platform (QGIS): Information on relative chronology is used to create three different views (Phase -1, Phase 0, Phase 1); for each view different symbologies are used to describe each feature. This plan is composed by overlaying features from different drawings and each feature keeps track of the source.

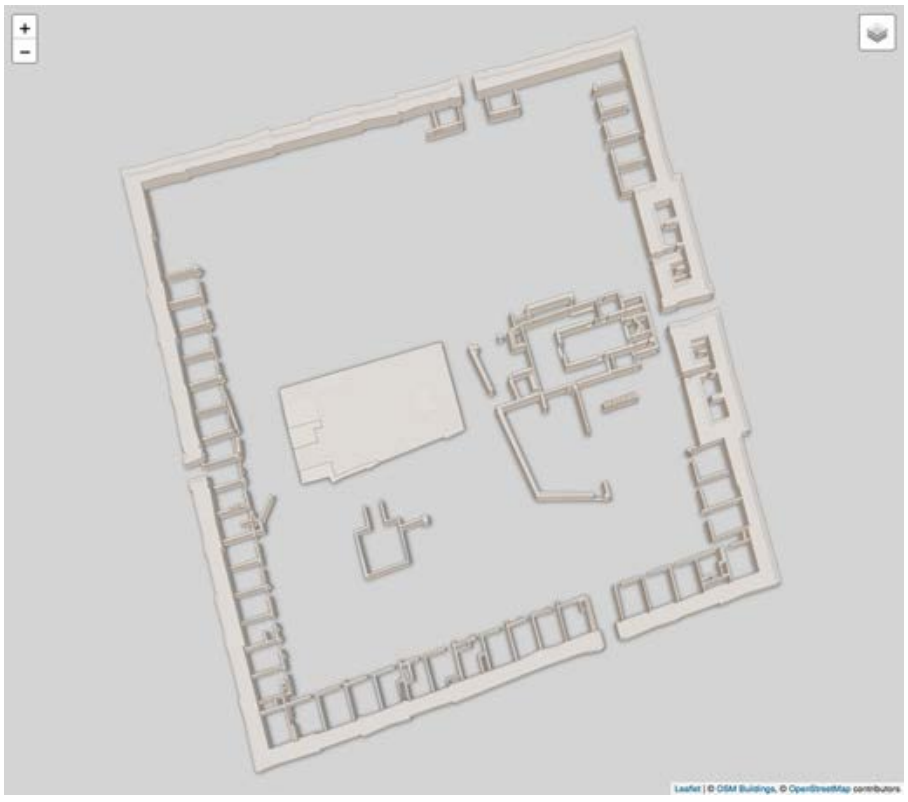

Fig. 5. Example of 2.5D (or pseudo 3D) view of the archaeological remains in the temple area of Taposiris Magna. The information for the extrusion of the structures is encoded using the SVP.

The only means we have to georeference these drawings is the satellite imagery freely provided as an XYZ service by Google or Bing.[46] In the most monumental cases, the archaeological remains visible in the satellite image allow the identification of discrete ground control point (GCP) fit to be used for the georeferencing process. It was decided to never deform the original image to better fit the geometry defined by the control points; not even in (the rare) cases when the identification of the GCPs was certain enough. Satellite imagery freely distributed by Google is nowadays an unavoidable source of information in the archaeological cultural heritage field,[47] yet the accuracy of these images is not defined, as their production does not rely on photogrammetric techniques. Empirical tests have pointed out that the error may differ by more than ten meters even in areas of high-resolution coverage.[48] The same conclusions might be observed by visually comparing satellite coverage distributed by different providers, such as Google and Bing.

For this reason, the relative metric references that the drawings contained (typically the scale bar and the north indication) were taken in a greater consideration than the visible remains on the satellite imagery. Furthermore, by not deforming the legacy images we acknowledge their historical value, which goes far beyond the metric precision of topographical survey. We, therefore, opted to process the published drawings in the same way historical maps have been processed in GIS environments. By now about 150 drawings of various size and complexity have already been georeferenced, but only the relative metadata have been published online, due to copyright restrictions.[49]

2.2.2 PAThs' Simple Vectorisation Protocol: towards a unified post-processing methodology for legacy data

The georeferencing process guarantees, on one hand, a very strict adhesion to the original data, but does not permit, on the other, efficient reuse of the data and their combination to produce a more coherent and updated graphical documentation, by combining, for example, multiple sources of information. For this reason, it was decided to vectorise the entire dataset to obtain a uniform, consistent and mixable data, that could be easily reused, for example in the online atlas.

The need for uniformity, consistency, extensibility, agility, versioning, and a high degree of abstractness lead to the drafting of a general-purpose protocol, for vectorising archaeological and architectonic drawings, named 'PAThs' Simple Vectorisation Protocol (SVP). It has been fully documented[50] and already tested on many tens of extremely simple to very complex cases. The SVP does not require any specific file-format,[51] rather it defines data structure and some best practices for the digital encoding of architectonical documentation in vector formats. The protocol offers an easy way to encode information about localisation of each element, its conservation status, a generic interpretation, relative chronology, metadata about the vectorisation process, the source of the information, etc., and essential information regarding the elevation. It consists of a set of attributes that can be added to a vector GIS file and few vocabularies to guarantee a uniform encoding of different sources of information.

The protocol is not intended to create perfect data but perfectible ones, i.e. a progressive approach is possible. For example, although one might have not information about the relative chronology of a building it is still possible to digitise the original drawing and keep the several phases separate (despite the fact that the order of the different phases is not known). At a later time, and after a careful study, the phase information can be easily updated, without affecting the geometries, to match a more coherent chronological seriation.

46 For a description of XYZ and other similar protocols for geographical data distribution in the Web, see Bogdani 2019a, 101-102.

47 Luo *et al.* 2018.

48 https://web.archive.org/web/20190716210821/http://www.uni-koeln.de/~al001/airphotose_files/hs1477.htm.

49 https://docs.paths-erc.eu/data/.

50 https://docs.paths-erc.eu/data/svp.

51 Any GIS vector file with support for attributes can be used with the protocol. At 'PAThs' both Shapefiles, https://www.esri.com/library/whitepapers/pdfs/shapefile.pdf, and GeoJSON, https://geojson.org, is being used, the later mainly for backup and web usage purposes. Moreover, since GeoJSON is a plain text file, it can be easily versioned using a version-control system, like Git https://git-scm.com.

Attribute name	Notes
place	In 'PAThs', the attribute is populated with the ID of the site where the building is found, eg. 338 for Taposiris Magna
subplace	Conventional name of the building as known in the available bibliography, if any, eg. Temple Church for the church located inside the *temenos*.
reconstr	0 (or *null*), 1 or 2. Set to 1 when the structure was not seen by the surveyor otherwise 0 or *null*. Reconstructed structures are usually rendered in publications with a dashed or dotted line. 2 is more rarely used in case the structure was not seen or hypothesised by the surveyor but was supposed by the digitiser.
part	s (or *null*), u, l, d. Describes the relationship of the represented element to the ideal section plain used to draw it. s (or *null*) stands for 'sectioned'; these are structures that are ideally cut by the section plane and are usually rendered with a hatch of oblique parallel lines u stands for 'upper projection' and is used for elements preserved above the section plan, such as covering, vaults, ceilings, capitals, etc. l stands for 'lower projection' and is used for elements preserved below the section plan, such as bases, altars, floor and floor decorations, benches, etc. d stands for 'doorways'; this is the only exception of the general rule of not encoding function because it would require interpretation and it would be prone to errors. Passageways and doorways are, however, and usually, sufficiently clear and easy to recognise. In uncertain cases, the more neutral l must be used.
phase	*null* or any positive or negative integer It provides a way to encode relative chronology when the original sketch differentiates phases, using colours, different strokes or hatches. Phases must be numbered from the more ancient to the most recent, although this information is not usually rendered in the sketch. If the information is missing, it is still necessary to differentiate the phases even if their order might not correspond to absolute chronology.
lost	0 (*null*) or 1 Defines whether a feature is visible (1) or not (0 or null). This attribute might be useful to update the state of preservation reported in the original publication. A feature might have been well-preserved whet it was surveyed (reconstr = null|0) but no more visible today (lost = 1) because buried or damaged.
scale	The scale of the original drawing from which the feature has been vectorised, if available. Enter only the second part of the ratio scale, presuming that the first part is always 1.
source	It contains a reference to the bibliographic or archive item used as the source for the vectorisation. At PAThs, the Zotero identifier of PAThs official bibliographic repository is being used.
subsource	Possibly page number and/or figure number of the bibliographic record containing the original image
operator	Name or codename of the operator who vectorised the image
date	Date and time of the vectorisation process
height	Elevation in meters of the feature, if known to use as extrusion parameter for 2.5D representation
minHeight	Offset elevation in meters of the feature, if known to use as extrusion parameter for 2.5D representation

Table 1. Attribute list of version 1.0.0 of the SVP, source: https://docs.paths-erc.eu/data/svp.

The SVP has proved to be a very flexible and powerful methodology because it allows to digitally encode in vector format almost any kind of architectonic drawing sticking very close to the original source and virtually with no loss of information. For the recovery of legacy data, this was a very important prerequisite. On the other hand, it offers also the possibility to emend old data, to update and to freely remix information from different sources and keep the entire process recognisable and fully reversible. Digital technology, and particularly GIS vector data allow us today to pursue efficiently the double goal of preserving the informative integrity of legacy data and expanding it with the potentiality of contextual and diachronic analysis. The Simple Vectorisation Protocol is nothing else but an attempt to abstract to a higher degree and at the same time make it easier to use the process of the vectorisation of legacy data, a process in which archaeologists have been involved for many decades.

Conclusions

The *Archaeological Atlas of Coptic Literature* is still under a very active phase of development and implementation and the 'PAThs' Simple Vectorisation Protocol, although it has proved to be an easy to use and powerful tool for the recovery of precious architectonic and archaeological data, is still little more than

an experiment. Nevertheless, the opportunity to work with legacy data, in the various meanings of this expression, has matured the realisation that digital archives are fragile containers. They can bring speed, efficiency, great user experience and almost infinite ways of connecting information, a decisive step in creating knowledge. They help us to manage and analyse massive amounts of information and they are no more an option. Yet our discipline has not matured a stable theoretical and practical framework to deal with them. Hesitant between an enthusiastic and messianic view and a more suspicious one, that considers digital technology as service that can be outsourced to IT engineers, we are still seeking for an equilibrium. Surely, we need to mature the awareness that the responsibility for the unparalleled quantity of digital data that we are collecting and partially publishing in this wonderful and highly volatile thing that is called Web cannot be delegated anymore.

References

Allison 2008 = P.M. Allison, "Dealing with Legacy Data – an Introduction", *Internet Archaeology*, 24, 2008 https://doi.org/10.11141/ia.24.8.

Barrett 2001 = J.C. Barrett, "Agency, the Duality of Structure, and the Problem of the Archaeological Record", in I. Hodder (ed.), *Archaeological Theory Today*, Cambridge, 2001, pp. 141-164.

Buzi - Berno - Bogdani 2018 = P. Buzi - F. Berno - J. Bogdani, "The "PAThs" Project: An Effort to Represent the Physical Dimension of Coptic Literary Production (Third-Eleventh Centuries)", in *Comparative Oriental Manuscript Studies Bulletin*, 4, 2018, pp. 47-66.

Bertoldi - Fronza - Valenti 2015 = S. Bertoldi - V. Fronza - M. Valenti, "Sistemi digitali di documentazione e analisi archeologica. Verso quale direzione", *Archeologia e Calcolatori*, 26, 2015, pp. 233-243.

Bloch 1953 = M. Bloch, *The Historian's Craft*, New York, 1953.

Bogdani 2016 = J. Bogdani, "Un archivio digitale multidisciplinare per la gestione e la conservazione di un patrimonio culturale a rischio: il progetto Ghazni (Afghanistan)", *Archeologia e Calcolatori*, Supplement, 2016, pp. 236-245.

Bogdani 2017 = J. Bogdani, "The Archaeological Atlas of Coptic Literature. A Question of Method", *Vicino Oriente*, 21, 2017, pp. 59-69.

Bogdani 2018 = J. Bogdani, "Linking Coptic Literary Manuscripts to the Archaeological Context by Means of Digital Humanities: the Case of the 'PAThs' Project", *Adamantius*, 24, 2018, pp. 200-210.

Bogdani 2019 = J. Bogdani, *Archeologia e tecnologie di rete. Metodi, strumenti e risorse digitali* (Cardini di Groma, 4), Roma, 2019.

Boyd - Lester 1974 = J.P. Boyd - R.W. Lester (eds.), *The Papers of Thomas Jefferson, vol. 19, 24 January to 31 March 1791*, Princeton, 1974.

Buzi 2017 = P. Buzi, "Tracking Papyrus and Parchment Paths: An Archaeological Atlas of Coptic Literature. Literary Texts in Their Geographical Context. Production, Copying, Usage, Dissemination and Storage ("PAThs")", *Early Christianity* 8, 2017, pp. 507-516.

Buzi - Bogdani - Carlig - Giorda - Soldati 2017 = P. Buzi - J. Bogdani - N. Carlig - M.C. Giorda - A. Soldati, "'Tracking Papyrus and Parchment Paths': A New International Project on Coptic Literature", *Rivista del Museo Egizio* 1, 2017, https://rivista.museoegizio.it/article/tracking-papyrus-and-parchment-paths-an-archaeological-atlas-of-coptic-literature-literary-texts-in-their-geographical-context-production-copying-usage-dissemination-and-storage/.

Cohen - Rosenzweig 2006 = D.J. Cohen - R. Rosenzweig, *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web*, Philadelphia, 2006.

Depauw - Gheldof 2014 = M. Depauw - T. Gheldof, "Trismegistos. An Interdisciplinary Platform for Ancient World Texts and Related Information", in Ł. Boloikowski - V. Casanova - P. Goodale - N. Houssos - P. Manghi - J. Schirrwagen, *Theory and Practice of Digital Libraries*, 2014, pp. 40-52.

Dunning 2001 = A. Dunning, "Excavating Data: Retrieving the Newham Archive", http://www.ahds.ac.uk/creating/case-studies/newham/index.htm.

Ellis 2016 = S. Ellis, "Are We Ready for New (Digital) Ways to Record Archaeological Fieldwork? A Case Study from Pompeii", in E.A. Walcek - J.M. Gordon - D.B. Counts (eds.), *Mobilizing the Past for a Digital Future*, Grand Forks, 2016.

Godlewska 1988 = A. Godlewska, *The Napoleonic Survey of Egypt: A Masterpiece of Cartographic Compilation and Early Nineteenth-Century Fieldwork* (Cartographica, 25, 1 & 2), Toronto, 1988.

Grossmann 2002 = P. Grossmann, *Christlicher Architektur in Ägypten* (Handbuch der Orientalistik, 62), Leiden - Boston - Köln, 2002.

Hodder 2005 = I. Hodder, *Theory and Practice in Archaeology* (Material Culture, 99), London, 2005.

Jablonka 2004 = P. Jablonka, "Reconstructing Sites and Archives: Information and Presentation Systems at Try", in K. Fischer Ausserer - W. Börner - M. Goriany - L. Karlhuber-Vöckl (eds.), *Enter the Past: The E-Way into the Four Dimensions and Cultural Heritage, CAA 2003, Computer Application and Quantitative Methods in Archaeology,* Proceeding of the 31st Conference, Vienna, Austria, April 2003 (British Archaeological Reports International Series, 1227), Oxford 2004, pp. 281-285.

Jacotin 1824 = P. Jacotin, "Mémoire sur la construction de la carte de l'Egypte", in C.L.F. Panckoucke (ed.), *Description de l'Egypte ou Recueil des observations et des recherches qui ont été faites en Egypte pendant l'expédition de l'armée française*. Seconde édition tome dix-septième. État moderne, Paris, 1824.

Jeffrey 2012 = S. Jeffrey, "A New Digital Dark Age? Collaborative Web Tools, Social Media and Long-Term Preservation", *World Archaeology*, 44.4, 2012, pp. 553-570.

Laužikas 2009 = R. Laužikas, "Digitization as Science", *Archeologia e Calcolatori*, 20, 2009, pp. 247-259.

Luo *et al.* 2018 = L. Luo, X. Wang, H. Guo, R. Lasaponara, P. Shi, N. Bachagha, L. Li *et al.*, "Google Earth as a Powerful Tool for Archaeological and Cultural Heritage Applications. A Review", *Remote Sensing*, 10.10, 2018, https://doi.org/10.3390/rs10101558.

Moreland 2006 = J. Moreland, "Archaeology and Texts: Subservience or Enlightenment", *Annual Review of Anthropology*, 35.1, pp. 2006, 135-151.

Moreland 2007 = J. Moreland, *Archaeology and Text* (Duckworth Debates in Archaeology), London, 2007.

Moshenska 2017a = G. Moshenska, "Introduction: Public Archaeology as Practice and Scholarship Where Archaeology Meets the World", in G. Mashenska (ed.), *Key Concepts in Public Archaeology*, London, 2017, pp. 1-13.

Moshenska 2017b = G. Moshenska, *Key Concepts in Public Archaeology*, London, 2017.

Orlandi 2003 = T. Orlandi, "Modeling the Coptic Literature. The 'Corpus dei Manoscritti Copti Letterari'", in *I&E 2002: Proceedings of the XIV Table Ronde Informatique et Egyptologie*, Pisa, 2003, http://www.cmcl.it/~orlandi/pubbli/pisa.doc.

Orlandi 2008 = T. Orlandi, *Coptic Texts Relating to the Virgin Mary*, Roma, 2008.

Previtali - Valente 2019 = M. Previtali - R. Valente, "Archaeological Documentation and Data Sharing: Digital Surveying and Open Data Approach Applied to Archaeological Fieldworks", *Virtual Archaeology Review* 10(20), 2019, pp. 17-27.

Rumsey - Williams 2002 = D. Ramsey - M. Williams, "Historical Maps in GIS", in K.A. Knowles (ed.), *Past Time, Past Place: GIS for History*, Redlands, California, 2002, pp. 1-18.

Valenti 2000 = M. Valenti, "La piattaforma GIS dello scavo. Filosofia di lavoro e provocazioni, modello dei dati e 'soluzione GIS'", *Archeologia e Calcolatori*, 21, 2000, pp. 93-109.

Valenti 2010 = M. Valenti, "Per un approccio neo processualista al dato archeologico", in A. Forgione - F. Redi (eds.), *VI Congresso nazionale di archeologia medievale* (*L'Aquila 2012*), Firenze, 2010, pp. 8-11.