Archaeological contents: from open access to open data

Aurélie G. Monteil, Viviane Boulétreau

HAL Id: halshs-01308043
https://shs.hal.science/halshs-01308043v2

Submitted on 13 Jan 2017

Archaeological contents:
from open access
to
open data[1]

Aurélie Monteil

aurelie.monteil@persee.fr

Viviane Boulétreau

viviane.bouletreau@persee.fr

UMS 3602 *Persée*. École Normale Supérieure, Lyon

Abstract:

The activity of the researcher takes place within a global process of development of the scientific knowledge, which relies on data sharing. He or she needs both to access existing data (sources, publications), and to distribute the material produced. Open access provides a suitable tool for reaching these two objectives.

Researchers must also overcome new challenges: selecting relevant information within the huge quantity of data and ensure optimal visibility and re-usability of the data they produce. In this context, open access is not sufficient, hence the emergence of the concept of open data, which deals with technical interoperability, quality of referencing and permanence of access.

Persée is a French, publically funded, program for the digitization and online publishing of printed academic journals in the field of Humanities and Social Sciences. Beyond the publication of a huge digital collection, Persée endorses the core principles of the open data movement (open access, handle/DOI, standard formats, SEO, interoperability, long term preservation).

The original mission of Persée is now expanded to a new project: producing and disseminating scientific content beyond published journals. Persée will apply its knowledge to a corpus composed of heterogeneous material. A wide range of tools is being developed in order to process, disseminate, share, and allow scientific uses of these bodies of documents.

Among the collections that will be processed with these new tools in 2015, two are dedicated to archaeology:

• The collection 'Monuments of Cairo' is about the digital publication of the minutes and reports of the Committee for the Conservation of the Monuments of Arab Art (published from 1882 to 1953). We intend to enrich the original material with a multilingual index for toponyms and monuments. This index uses international standards and proposes to associate to each entry the several ways it is mentioned over the whole collection.

• The collection relating to the excavation of 'Salamis of Cyprus' concerns the study of material and architectural remains. It aims at firmly identifying artefacts and establishing links between several resources related to these objects: catalogues, photographs, publications, index cards.

These two projects will demonstrate the potential brought by open data regarding the constitution of digital collections. One of our objectives is to build a large-scale platform that will both federate data and be enriched, project after project, by a collection of tools addressing researchers' needs. Both data and tools will be fully and freely available to the community.

Key Words: *open access, open data, interoperability, XML, linked data*
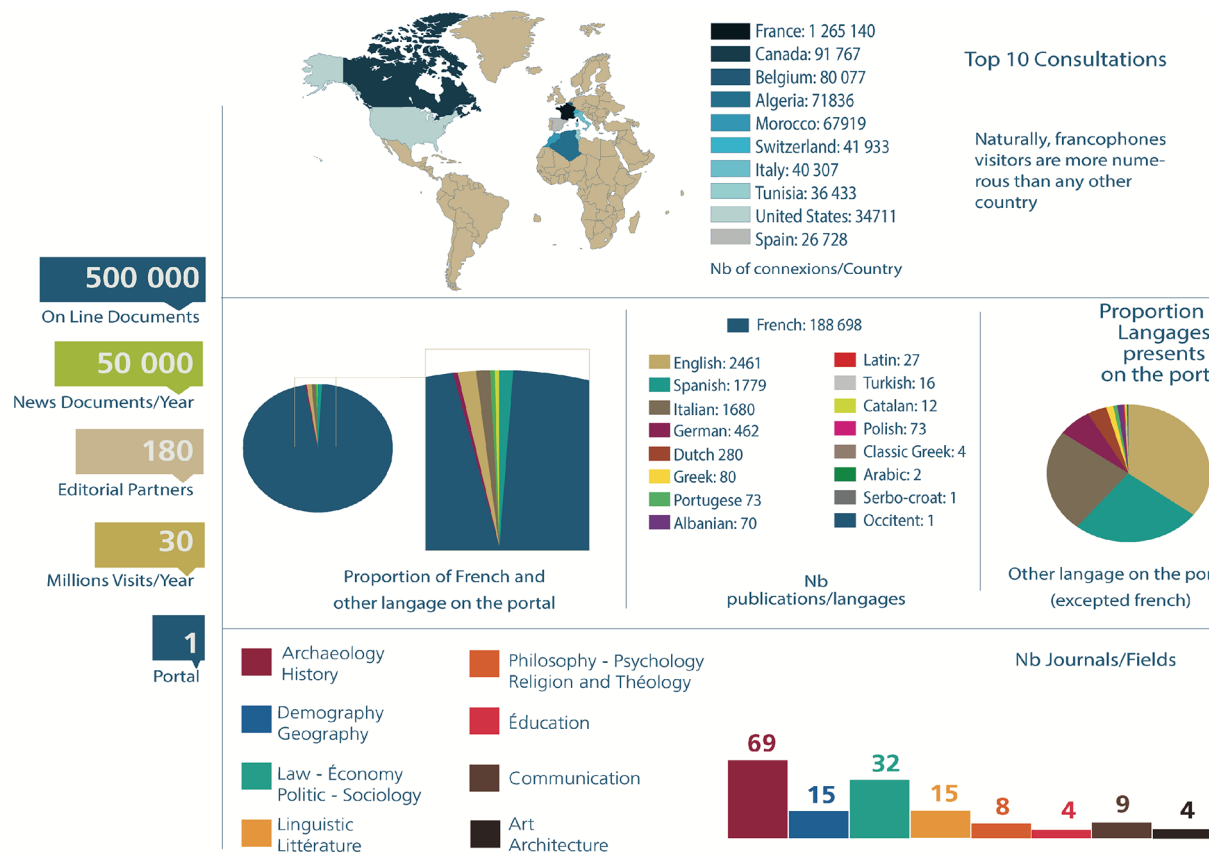

Fig. 1. Characterization of collections available on the *Persée* platform

Introduction: the French UMS Persée

Persée is a French public-funded program for the digitization and online publishing of printed academic journals in the field of Humanities and Social Sciences.

In 2003, the French minister of Research and High Education conducted a public call for proposal. The University Lumière Lyon 2 was selected to realize the *Persée* program.

Since 2012, *Persée* has developed new partnerships with the École Normale Supérieure de Lyon, which physically hosts the team, the French Center of Scientific Research (CNRS) and the University of Lyon. The program became an unit (UMS, Unité Mixte de Service), devoted to digitization, enrichment and dissemination of scientific documentation.

The goal set to the unit is to provide a support service for the development of scientific research. It aims to enhance the visibility of French scientific production with several key objectives:

• promote free access

• disseminate enriched data

• carry out an active policy of contents referencing within international systems (author databases, search engines, Cross-Referencing DOI)

• ensure the interoperability of data (OAI-PMH and Z3950)

• use open formats (METS > descriptive, administrative and structural metadata ; MODS, DC, MarcXML >bibliographic metadata ; MADS > authority metadata ; Erudit schema, TEI > full text)

• promote enriched uses (bibliographic tool, browser software, output/portable enriched format)

1. Collection Holdings

Our Journals collection is characterized by 3 particularities:

• Language: our collection is mainly composed of francophone journals, in its initial demand of the Ministry sought to promote the French scientific production,

• Scope: since the foundation of the program, our specialty has been the field of humanities and social science. Since the beginning of 2015, we have been completing our collection with other subjects such as earth and life sciences (collections in biology, mineralogy, botany, glaciology are currently being processed), medical sciences or engineering sciences and techniques.

• Selection: The last particularity is the selection of the journals we disseminate: they should respond to a high level of scientific standard, and must be listed by an Institutional Reference Index as: ERIH (the European Reference Index for the Humanities) or the AERES (the Evaluation Agency for Research and Higher Education). The selection of journals is achieved by a board of scientific examiners.
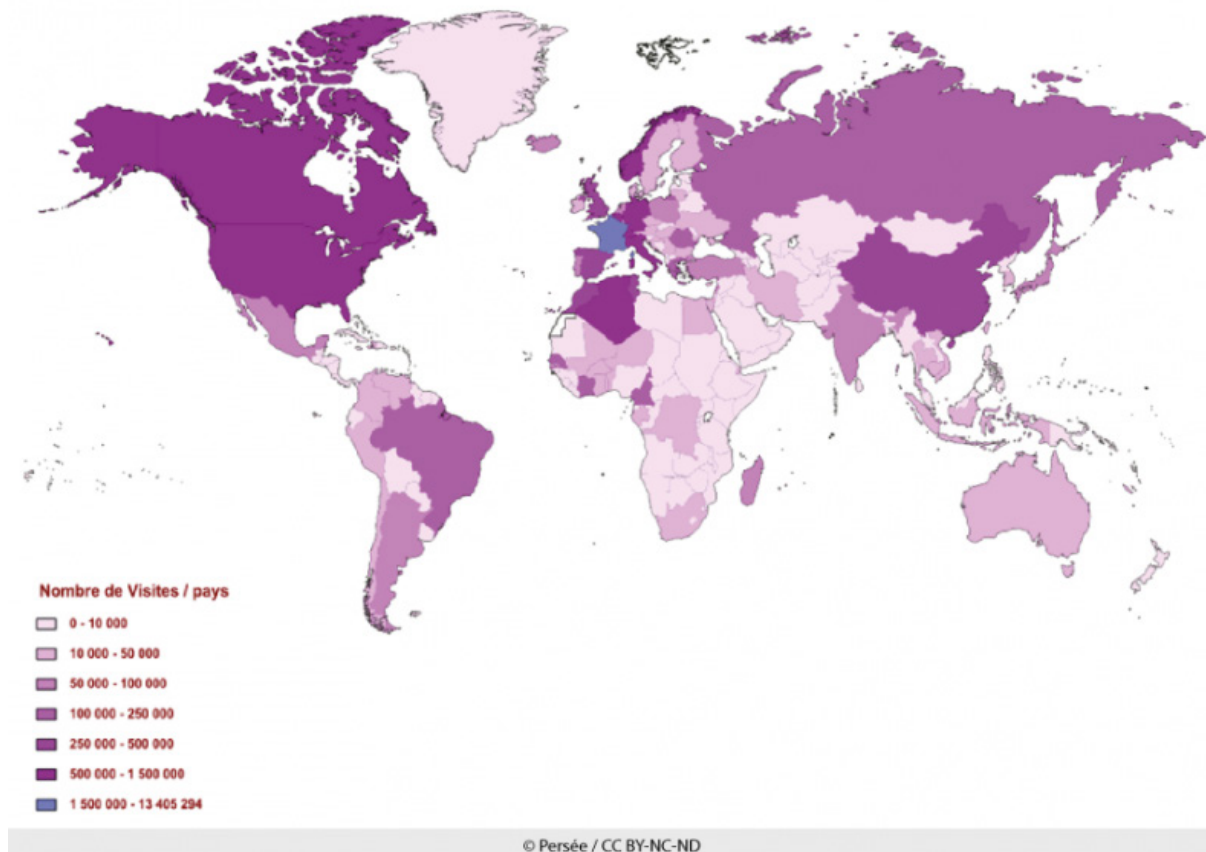
Fig. 2. Global distribution of visits to persee.fr

Once selected, the journal contracts a non-exclusive agreement with the unit, then the whole process of digitization and documentation is endorsed by *Persée*.

With 559, 088 publications in French, this language is mostly representative because of our politicy concentrate on the promotion of French Journals. It is followed by English language (8,647 publications), then Spanish (2,491) and Italian (1,955).

Of course, this goes hand in hand with the leading connexions by country, which, unsurprisingly, come from francophone countries or with a well-established francophone culture (Algeria, Marocco).

2. Process line

Persée processes exhaustive collections, from the first published issue to the most recent ones. The process line has two entries: the first process is based on paper journal. All books are cut and therefore the treatment is destructive for books. The second entry is based on the processing of digital documents like PDFs. From both entries, we produce images and texts in XML format and process data with our own application. Then the document is enriched with a bibliographical metadata record and an internal structure description, i.e. the titles, footnotes, bibliographic areas, illustrations, etc., are fully described.

The process line leads to a complete set of data allowing us to reach our objectives:

• upload the *Persée* portal with complete numeric collections

• allow rich uses of documents (browse, query, cite, …),

• facilitate data sharing (use of standard formats)

• ensure the preservation of data by archiving them on a long term preservation platform

3. *Persée*'s main objective: reaching Open Data?

What are the main principles of Open Data? The best answer is given by the "BBB definition", this acronym stands for 3 declarations known as Budapest (2002), Bethesda (2003) and Berlin (2003) (The BBB Definition is identical to the definition of the Budapest Open Access Initiative, see Budapest 2002). The "Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities" (Berlin 2003) written on October 22, 2003 defined Open Access as: 'a comprehensive source of human knowledge and cultural heritage that has been approved by the scientific community'. This declaration is completed by this consideration: 'In order to realize the vision of a global and accessible representation of knowledge, the future Web has to be sustainable, interactive, and transparent. Content and software tools must be openly accessible and compatible.'
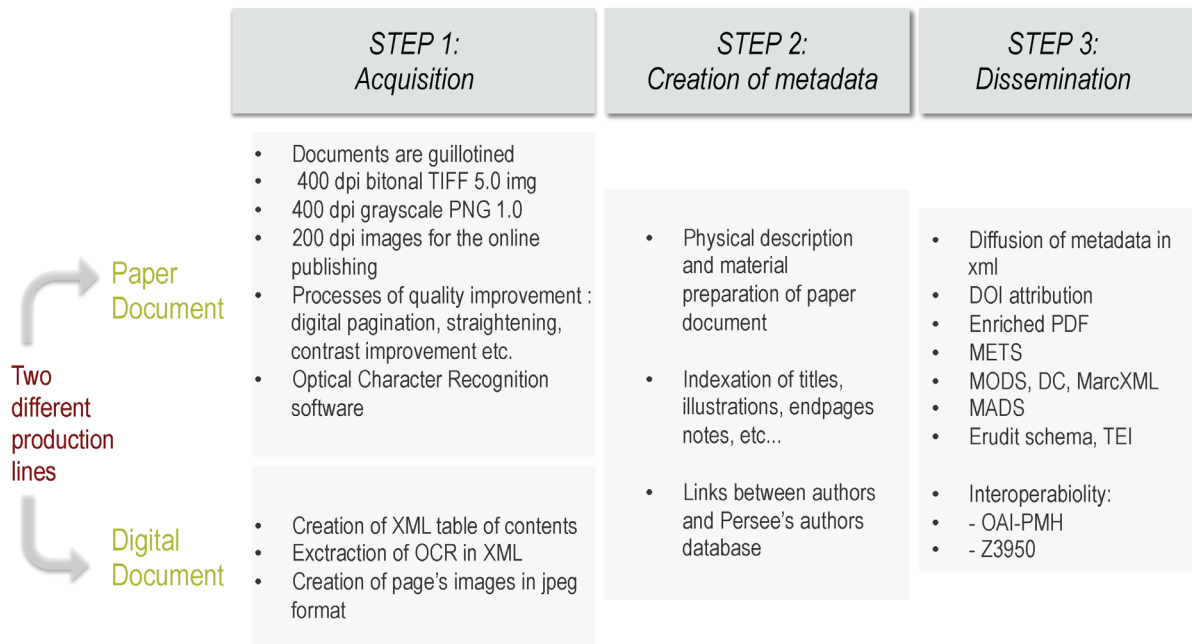
STEP 1: Acquisition	STEP 2: Creation of metadata	STEP 3: Dissemination
Paper Document • Documents are guillotined • 400 dpi bitonal TIFF 5.0 img • 400 dpi grayscale PNG 1.0 • 200 dpi images for the online publishing • Processes of quality improvement : digital pagination, straightening, contrast improvement etc. • Optical Character Recognition software	• Physical description and material preparation of paper document • Indexation of titles, illustrations, endpages notes, etc... • Links between authors and Persee's authors database	• Diffusion of metadata in xml • DOI attribution • Enriched PDF • METS • MODS, DC, MarcXML • MADS • Erudit schema, TEI • Interoperabiolity: • - OAI-PMH • - Z3950
Digital Document • Creation of XML table of contents • Exctraction of OCR in XML • Creation of page's images in jpeg format		

Two different production lines

FIG. 3. GRAPHICAL OVERVIEW OF THE *PERSÉE* PROCESS LINE SHOWING THE MOST IMPORTANT STEPS

In order to success in an open access dissemination of data, several main criteria must be validated: the data must be correctly generated in order to be understandable and re-usable, their localization must be stable and known, tools for the management of their update and sharing must be settled up. For each of these points, we will mention the "bad way", give main keys for the "good way" and illustrate our point with examples from the world of XML and archaeology.

3.1 The representation of data

When creating digital data, searchers are mainly concern with the use they plan to make of it. The use others may develop are not envisaged. This leads to databases produced with very specific models, XML documents based on local (mostly unpublished) schemas, indexes built upon personal vocabularies, etc. Hence, most of the time, the way the producer represents its data is not documented. Such practices make it difficult, even impossible, the use of data outside their original creation context.

Adopting two main good practices can ensure the future usability of data.

First, whatever the involved technology may be (SQL, XML,…) the choice of the model on which the representation of data will be based is a main concern. The producer should favour a standard one.

In the XML world, two schema are widely used for the representation of data: the Text Encoding Initiative (TEI), devoted to the representation of texts, their annotation, interpretation, etc., and the Encoded Archival Description (EAD) which is inspired from TEI, but is devoted to archival material (images, objects, etc.). Both of these models are very rich, fully documented and have a large community of users. The fullness of these models may alarm the searcher when he is not familiar with them, but in most cases, he or she will not have to deal with the whole complexity of the model but

will focus on a subset of it, making obsolete the need for a complete understanding of the model refinements.

The application field of this first principle should not be limited to data, metadata must also be considered. All the knowledge that go with the main resource should lay on the use of standards: the Dublin Core for documentary description, international identifiers for persons (ISNI[2]) or places (Geonames[3]), and referencing vocabularies each time one is available.

The second good practice is even less well respected by data producers: it is the documentation of the data creation process. Since the production of data is mostly punctual, associated to a particular project, funding or researcher, the particular scientific context in which they are conceived necessarily leads to choices in the way data are represented.

3.2 The dissemination of data

The way data is disseminated should also be a main concern for researchers. If it is convenient to hold them on the lab web site, it is not enough to make them trustfully usable in a medium-term perspective. The system hosting the site may change, the researcher who manage it may leave, etc. The sustainability of the solution must be considered. The dissemination solution must give some guarantees in terms of accessibility, stability and sustainability. It should also endorse an active policy of web referencing, in order to ensure the visibility and the promotion of data.

3.3 The identification of data

Once data are accessible, there is one additional condition

---

[2] International Standard Name Identifier (ISO 27729), http://www.isni.org/

[3] GeoNames geographical database, http://www.geonames.org/

for their re-use by other systems; it is the trustworthiness that one can grant to its accessibility. No one would spend time to base work on data that is susceptible to move, disappear, etc., without offering any warranty of constancy. The existence of an identification of a resource, in the intellectual sense, regardless of its current location of the web, is the only way to preserve the user (data provider or re-user) from broken links. The attribution of such identifiers should be ensured by the producer or by the system on which the dissemination is based. In the academic world, two main international systems are available to identify resources: DOI maintained by CrossRef[4] is devoted to published material; handle.net[5] can be used for any type of content.

3.4 The preservation of data

Of course, a protocol for the Long Term Preservation of data (storage and format) is welcome. This main concern is not limited to data itself, but also to the technology suited to their use. The norm proposed by the Open Archival Information System (ISO-14721[6]) defines several requirements in order to guaranty the preservation of digital material.

Their implementation is quite heavy, but in each country, several institutions have settled up such archives and propose to host data.

3.5 The mutualisation of data

Last, but not least, the system chosen for the dissemination should not be limited to a web interface, but should also provide standard gateways to communicate with other platforms, both for the initial sharing of data (harvesting), and for the management of their eventual updates.

The system should also provide interfaces allowing a true exploration and exploitation of data by other programs (Application Programming Interface, widgets, etc.). This particular facet of data sharing is called interoperability. It allows anyone to re-use data with approaches or objectives different from the original ones.

3.6 Rights management

Openness has two facets: technical and legal. Making data available in a way that technically suits their use by other researchers is not sufficient in itself – the user should also be granted with the adequate rights to use them. Many licences exist to define which the user's rights regarding data are. In the scientific world, the most commonly used are 'Creative Common [1]' licences, which offer a wide variety of solutions, from very restrictive licences to public domain ones. All 'Creative Common' licences are ruled by common conditions: to provide a non-exclusive licence to produce, distribute and transmit the data to the public freely; to respect the author copyright; to allow a peer to peer network.

Persée opted for the 'Attribution-NonCommercial- NoChange' licence (CC BY-NC). This allows redistribution in a

commercial and non-commercial way, as long as it is passed along unchanged and in whole, with credit to its original provider (https://creativecommons.org/examples/).

In summary, *Persée* only authorizes the reproduction and distribution of the original work, this licence requires those using data to credit us for our work, and not to suggest that we approve its use or give our endorsement or support, and finally, we allow others to copy/distribute without changes. We refer to these conditions on the first page of all our downloadable documents.

4. Implementation of these principles

The advantages of using Open Access are obvious: the main argument is the free licensing software, which is an essential point for academics. Then, thanks to free sharing, we can assure the durability of data thanks to a huge community of users. The Open Data is necessarily commutative and thus collaborative. The huge and free dissemination of scientific knowledge is a necessary way for the scientific progress. Now, we know what the principles and the benefits of Open access sharing are. To implement an active policy of Open protocol, one should apply several uses in terms of technical competencies and applications.

The researcher and the scientific institution must set up a technical infrastructure to assume open data diffusion. Most of the time, this economic investment is a problem for a laboratory, in this case the researcher can make the choice to use an external open access platform so as to disseminate his work, i.e. the French HAL or Academia.com, ResearchGate, etc. Whatever the dissemination method you choose, the open data corresponds to a community of customs and uses. Among these different formats, *Persée* has chosen the most frequently used. For sharing data, we must keep in mind the three types of existing exchanges:

- the users looking for information

- the computer harvesting or querying the data repository

- and the data transfer to preserve.

The 'users' must be guaranteed by the free and unlimited access and the consultation of the document in full text without limitation. The 'computer' must allow the collection of information using standard formats. Specifically in the case of the *Persée* program, this approach is reflected by the selection of the formats we have implemented:

- Data are represented both according to the TEI[7] and the Erudit[8] schemas

- Metadata are stored in Dublin Core, XMLMarc and MODS[9]

- Authority records are in MADS[10]

- The organization of data (structure of issues, of collections, etc.) is stored according to the METS[11] schema

4 CrossRef: http://www.crossref.org/
5 Handle.net: http://www.handle.net/
6 OAIS, ISO-14721 norm, http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284

7 Text Encoding Initiative: http://www.tei-c.org/index.xml
8 Erudit schema: https://apropos.erudit.org/en/erudit-en/technologies/
9 Metadata Object Description Schema (MODS): http://www.loc.gov/standards/mods/
10 Metadata Authority Description Schema (MADS): http://www.loc.gov/standards/mads/
11 Metadata Encoding and Transmission Standard (METS): http://

The services associated to data are numerous:

• In order to identify resources, we attribute DOI to each of the articles available on the *Persée* platform

• Procedures for the long term preservation are integrated in our process line. The archiving is realized by the CINES (French national platform for higher education and research)

• An OAI-PMH[12] server allows automated harvesting of metadata and full-text from any archive repository

• A Z39.50[13] gateway is up for the interoperability with library catalogs

• Soon, a triple store will be open allowing searchers to perform specific and/or complex query on *Persée*'s collections

• Collaborations have been established with the main editors of discovery tools in order to include the *Persée* contents in their tools

• Technical choices and procedures have been described and published in order to be shared with the community. For example, our guidelines have already been taken up by French digital libraries as Revue.org and Cairn

• For the portal user, mechanisms allowing him to feed its bibliographic tools are available (EndNotes, Zotero, etc.)

• Authors are identified and linked to the referentials managed by the national bibliographic agency, the national library, Wikipedia, etc.

• Citations are identified and declared to CrossRef

• Up to now the Persée data and metadata are disseminated with quite restrictive licences: CC-BY-NC-ND. Users are allowed to use them in a non-commercial context, have to mention the origin of the data, and are not allowed to distribute any derived form of the content.

• The Persée team advocates the adoption of more open licensing for metadata and 'old publications'. The scientific board of the unit should decide (Spring 2016) whether this material can be disseminated on a CC-BY-SA or ODbL13 licence, which would allow any user to do anything (including commercial uses) with the data provided they include a mention of the data's origin and keeps the core data open.

5. Persée's Project

Since 2003, this reflection has underlied our work for scientific Journals. We now apply all these principles to new kinds of materials: 'corpus' characterized both by the heterogeneity of the documents and by the specialized tools to be settled up in order to satisfy the searcher's needs. Following are two cases which illustrate these points, two projects in the fields of archaeology[14]. The objectives of these new projects are the processing of scientific corpus, the digitization of exhaustive documentation and the input of an added value with augmented metadata:

• the first project is the 'Athar Project' - on Cairo ancient Monuments in Egypt.

• the second project is on 'Salamis (Cyprus)' - concerning the complete archaeological excavation documentation.

Both projects are based on the digitization by *Persée* of a printed collection, but each one is enriched involving different approaches. For Salamis the printed material constitutes the 'global information' from which a refined indexing of smaller objects has been achieved. For Athar, it is precisely the opposite approach: the printed material is the refined material. Associated with the toponyms, it becomes substantially extended by knowledge of various factors such as a SKOS toponymic and multi-linguistic index, the geolocation of the monument, etc.

In the next section, we will describe how we have decided to structure this documentation, how to disseminate it and how these choices lead to a considerable added value.

5.1 The Athar Project

The project "Athar" has for ambition to digitize and make available the works of the Committee for the Conservation of the Monuments of Arab Art. This yearly publication identifies, lists and describes all ancient Islamic and Copt monuments in order underpin the management of their restoration. For the historian or art historian who studies the historiography of Islamic art, the history of preservation and the patrimonial monument of Egypt, this corpus constitutes a unique data source. Its value is based on the number of described monuments, the quality of architectural and historical information, the amount of technical documentation taken from restoration worksites, the prosopography of actors and all the photographic archive, etc. Its great scientific interest is based on the fact that so many ancient monuments have been destroyed (today on 800 monuments identified since 1880, 300 have already disappeared). The preservation and dissemination of these data are considered as important by the scientific community. Our project consists in disseminating this collection allowing a classical document display and an enriched TEI text for more refined searching. The main difficulties are the typological treatment for the transliteration of Arabic names. Our chief goal was the elaboration of methodological tools and the provision of a large index for the enhancement of this textual and iconographic corpus. This project illustrates the method used for transforming a collection, which is hard to access, into a rich set of information available in open way, giving a new life to this information.

The tool known as 'Jgalith' developed by the *Persée* team allows the creation of metadata. The first step for indexation is the insertion of the index of ancient monument names index built by the researchers themselves and provided by our partners. Once this index inserted in our information system,

---

www.loc.gov/standards/mets/
[12] Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH): https://www.openarchives.org/pmh/
[13] Z39.50 (ISO-23950): "Information Retrieval: Application Service Definition and Protocol Specification": http://www.loc.gov/z3950/agency/
[14] *Persée* has also selected projects in other scientific area: NOOEE:

is a corpus in biology. The goal is the digitization of publication or collections of specimen and the indexation of taxonomic name of species.
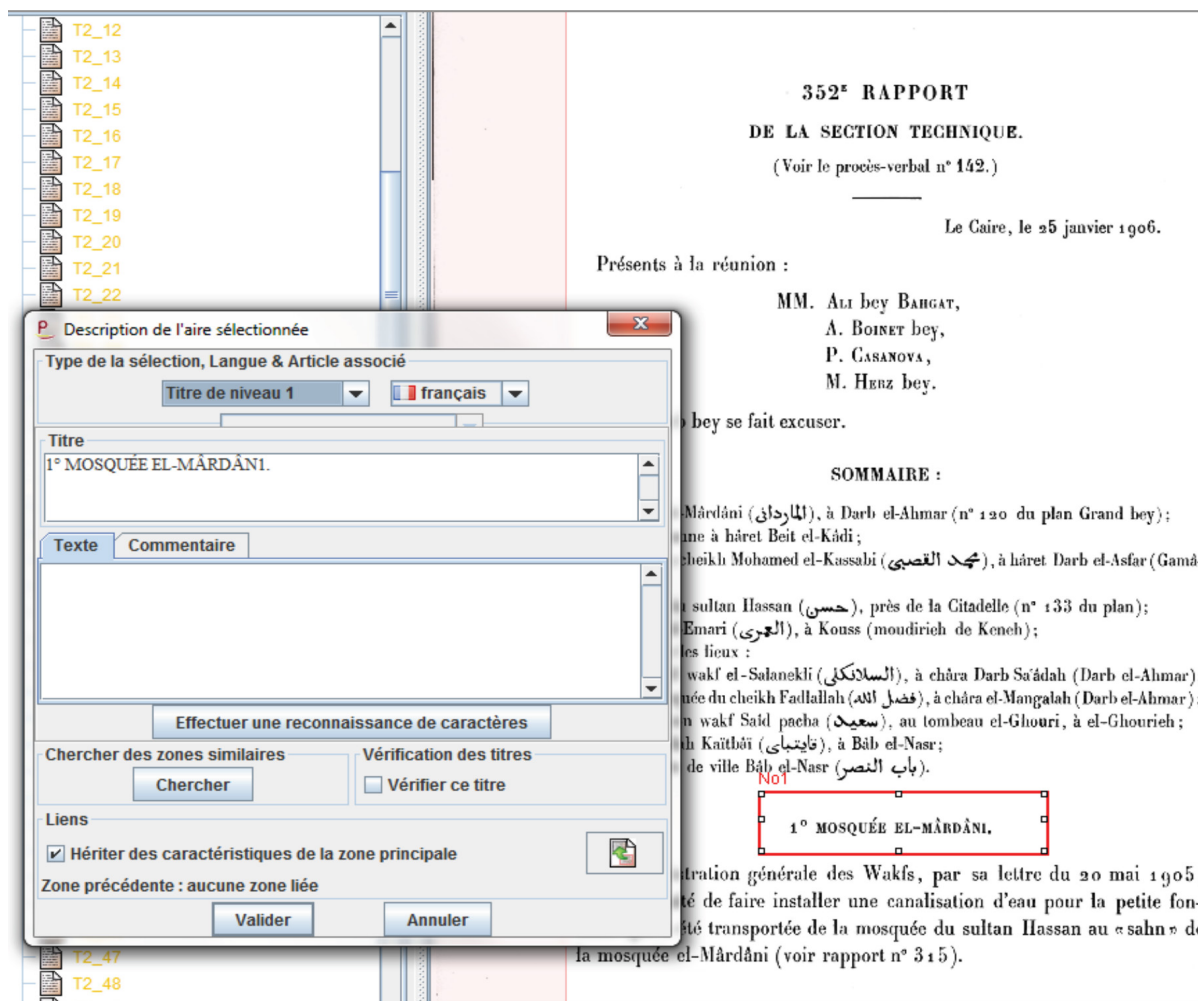
Fig 4. Exemple of the 'Jgalith ' Persée application.

it is first used as a supplementary dictionary which can be involved in the optical character recognition process and enhances the quality of the text produced from the analysis of images of pages. Then, when the monument name is recognized, a specification about its typographical form is given: Is it a cursive Arabic form? A transliteration? Once this work achieved, the multilingual Index with the name of all ancient monuments integrating all information about them and presenting all the existing typographic forms for its names will be spread. The final objectives are to propose a large set of XML (encoded in TEI) documents with tool for their indexation. The Corpus constitutes the global base for the dissemination of information. The name index has been inserted in the geocorpus Geonames and presents the site and its location within the city.

These information also feed other platforms:

• the MarcXML metadata is feeding the Sudoc and the library catalog.

• the Dublin Core and TEI are used for the interoperability with other Open Access platforms, like the Canadian one Érudit or the French platform Isidore for the open dissemination in Social Sciences and humanities. Last, a triplestore will be settled up with a SPARQL endpoint for querying.

5.2 The Salamis of Cyprus Project

The second project is the digitization of the archives of the French archaeological survey to Salamis (Cyprus). This corpus constitutes a vast and unique documentary collection about this ancient site. The excavations started in 1964 under the direction of Jean Pouilloux and have been continuing until 1974. At this date, half of the island became Turkish, and the access to Salamis became forbidden, all the work on the former city stopped at that time. From this date, the ancient site has remained inaccessible to research. Today, most of the artifacts are destroyed or have disappeared, making even more important the preservation of the remaining objects. This documentation constitutes the only available documentation on the archaeological site of Salamis. This project is of particular importance since it allows at the same time the scientific and patrimonial enhancement and ensures the preservation of the remaining documents. The digitization ensures a huge dissemination of these primary data, the development and renewal of research about this archaeological site and provides for the good conservation of the documentation. Some of the files are so old that they are starting to deteriorate; these older documents are fragile and the digitization would allow access without a risk of damage or destruction. Both partner institutions provide the

7

Fig. 5. Results of digitization and enrichment on persee.fr

production of data:

• The Maison de l'Orient et de la Méditerranée provides the digitization of the excavation documentation (field notebooks, excavation photographs, index cards, maps, drawings) which constitutes all the documentation relevant to an excavation. All documents and informations are represented with the XML MODS schema which allows the complete description of the various kinds of resources.

• *Persée* digitizes the scientific publications and creates metadata and bibliographical records, attributes Digital Object Identifier for each document, establishes links between the authors and the database IdRef, and produces all metadata usually going with serial publications.

When the production step is achieved, the dissemination of all data begins. Figure 5 illustrates the potential of interoperability between our data and the data provided by the Maison de l'Orient et de la Méditerranée. It presents a sketch of what the consultation interface may be for the Corpus of Salamis: a query within this corpus leads to parallel resources: on the left side, the content of the printed material is provided by *Persée* and the list of other material, with their identification numbers, produced by the archeologist during the excavation. Additionally, links between terms found in the text of the publication and the same terms in the index of the gray literature allow the provision of external links to several associated documents.

As a conclusion, it needs to be stressed that open data requires a strong application support, which supposes a technical team and thus, a sustainable and solid investment from institutional structures. Even if institutions develop policies of open access, there is no guarantee that these policies will be prolonged, which constitute a main difficulty. The best way for researchers is to resort to larger institution specialized in the management of Open Data. Today, the offer is wide and the selection of the appropriate solution should meet two main conditions:

- Human and technical support

- Political and economic stability

These two conditions are essential for the viability of a project. Today, relying on such institutions is the best way for a researcher to succeed in the good spreading of scientific data. The multiplication of open scientific platforms encourages the personal and voluntary action of open dissemination. This individual action must be supported by active policies from universities and research institutes.

The Budapest Open Access Initiative. [Online] http://www.budapestopenaccessinitiative.org/read [Accessed: 26th June 2015]

Berlin 2003
The Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. [Online] http://openaccess.mpg.de/Berlin-Declaration [Accessed: 26th June 2015]

Budapest 2012
Ten years on from the Budapest Open Access Initiative: setting the default to open.[Online] http://www.budapestopenaccessinitiative.org/boai-10-recommendations [Accessed: 26th June 2015]

Berners-Lee 2006.
Berners-Lee T. , *Linked Data - Design Issues*. [Online]. https://www.w3.org/DesignIssues/LinkedData.html

Bizer 2009
Bizer, C. The Emerging Web of Linked Data. In: *Intelligent Systems IEEE* 24, 5: 87–92. [Online] doi: 10.1109/MIS.2009.102

Bibliography

Budapest 2002