The Impact of FAIR Principles on Scientific Communities in (Digital) Humanities. An Example of French Research Consortia in Archaeology, Ethnology, Literature and Linguistics

Adeline Joffres, Nicolas Larrousse, Stéphane Pouyllau, Olivier Baude, Xavier Rodier, Michel Jacobson, Véronique Ginouvès, Fatiha Idmhand

The Impact of FAIR Principles on Scientific Communities in (Digital) Humanities. An Example of French Research Consortia in Archaeology, Ethnology, Literature and Linguistics.

Adeline Joffres (adeline.joffres@huma-num.fr), CNRS, Huma-Num, France

Nicolas Larrousse (nicolas.larrousse@huma-num.fr), CNRS, Huma-Num, France

Stéphane Pouyllau (stephane.pouyllau@huma-num.fr), CNRS, Huma-Num, France

Olivier Baude (olivier.baude@huma-num.fr), CNRS, Huma-Num, France

Fatiha Idmhand (fatihaidmhand@yahoo.es), Université de Poitiers, France

Xavier Rodier (xavier.rodier@univ-tours.fr), Université de Tours, France

Véronique Ginouvès (veronique.ginouves@univ-amu.fr), Université d'Aix-Marseille, MMSH/phonotèque, France

Michel Jacobson (michel.jacobson@huma-num.fr), France; CNRS, Huma-Num, France

The French TGIR Huma-Num, whose mission is to facilitate the digital turn in Humanities and Social Sciences research, offers services dedicated to the production and reuse of data. These services aim at avoiding loss and facilitating the reuse of scientific data in Social Sciences and Humanities. To do this, Huma-Num supports research teams and disciplinary consortia throughout their digital projects to allow the sharing, reuse and preservation of data thanks to a chain of devices focused on interoperability. Through these processes, Huma-Num also encourages compliance with the FAIR data principles.

Huma-Num's tools connect normalised metadata to the Linked Open Data cloud and give them an extended visibility. By labelling consortia (through scientific, financial and technical support when needed), Huma-Num drives and supports them in this initiative: data producers are encouraged to clean and normalise their data, they benefit from Huma-Num's services to store, share, expose, signal and enrich their data. What's more, at the end of the chain, these processes also allow the data to be "ready" or, at least, "prepared" for archiving, which is a real issue for research data.

But what is the effect induced on the data producers by sharing their resources, particularly on the metadata? Do they realize that the metadata produced for a specific project are not suitable for more generic needs and need some polishing?

Does this virtuous circle produce in turn better metadata and also better practices?

The panel will try to answer these questions by presenting feedback from different disciplines and communities in order to trigger discussion. More specifically, through the panel, four of Huma-Num's consortia - in archaeology, ethnology, literature and linguistics – will present their experience, practices and some tools they have produced in order to measure the impact of this supposedly virtuous circle on the quality of the data and metadata they have produced and exposed in the LOD. Additionally, this will allow us to discuss the role of a national research infrastructure like Huma-Num in France and the collaborative and expert networks developed through the creation of Huma-Num's consortia.

By sharing the points of view of various disciplinary but multi-approach consortia in Social Sciences and Humanities, the panel will aim at proposing a reflexive approach to the impact of the FAIR principles.

The panelists will present the status of the different subdomains with very brief talks (15 minutes each), highlighting opportunities, consolidated approaches and open issues. The different perspectives and experiences presented, on various types of data and in different disciplines, will build a common space for the further Q&A discussion with the audience on the application of the FAIR principles in the DH domain.

1. Panelists

- Michel Jacobson, CORLI consortium, CoCOon, https://cocoon.huma-num.fr/exist/crdo/
- Fatiha Idmhand, CAHIER consortium, CNRS/Huma-Num, http://cahier.hypotheses.org/
- Xavier Rodier, MASA consortium, CNRS/Huma-Num, https://masa.hypotheses.org/
- Véronique Ginouvès, AdE consortium, CNRS/Huma-Num, https://ethnologia.hypotheses.org/

2. Chairs

- Adeline JOFFRES, TGIR Huma-Num, CNRS, France
- Nicolas LARROUSSE, TGIR Huma-Num, CNRS, France

3. Panelists - Talks abstracts

3.1. Talk 1 by Véronique Ginouvès, MMSH/CNRS, "Archives des Ethnologues" Consortium, France.
Anthropologists, Archivists and Fieldwork Materials: Best Practices of a French Consortium

For the resource centers that compose it, the creation of the Consortium "Archives des ethnologues" within the Huma-Num TGIR in 2012 was key to the acquisition of new methodological reflexes in the digital domain. The aim of its eight partners is to store, process, share and publish the documents produced by anthropologists in their fields.

The scientific and heritage importance of these survey materials, the richness and diversity of the societies studied, oblige us to take the singularity of these data better into account.

We will show how the consortium is working towards greater convergence of the practices of description, structuring and access to ethnological data, notably through the FAIR Data principles ("findable, accessible, interoperable, reusable").

These practices give us the opportunity to adapt these data to documentary projects or scientific research and their objectives.

Thus, in order to enhance the discovery and availability of these archives, the metadata used to describe them are not only produced in DC (Dublin Core) but also, for example, in EDM (Europeana data model) or EAD (Encoding archival description).

Similarly, their transfer to different data archives (ODSAS, Kinsources, Portal of oral heritage, MediHal, ...) and their access via different platforms (Calames, Clarin, Europeana, Isidore) are essential for their availability, research, identification and increased interoperability.

We will also focus more specifically on the use of vocabularies, which are essential to improve the search for these data on major platforms.

The first example will be that of the authorities. We will show how the use of tools such as ISNI, VIAF or IDREF "propel" the members of the societies studied by ethnologists into international standard name systems. These witnesses, who have remained anonymous for a long time, then become truly authors of their word and their families can thus find traces of their names through the devices put in place.

The second example of a vocabulary that will be presented is the creation of a thesaurus (in SKOS format and produced with OpenTheso software) on uniform titles of tales in order to enrich the research on oral literature.

Finally, we will also address the issues of data access and reuse.

From the beginning of the Consortium, discussion has been ongoing to address the ethical and legal issues related to the dissemination of humanities and social sciences data. A blog (https://ethiquedroit.hypotheses.org) provides answers to the concrete questions that arise during online publication and a guide of good practices will be published in September 2018. We will also present some examples of the re-use, on our platforms, of some of the data processed by the different centers: a crowdsourcing project with Transcrire, a work space for research with ODSAS or the provision of specific data (the kinship data) with Kinsources.

These principles are also the strength of national institutions (TGIR Huma-Num, CINES) that support the implementation of projects over the very long term. They provide us with a solid framework for organizing data life, access and sharing.

3.2. Talk 2 by Xavier Rodier, UMR CITERES-LAT, Tours University/CNRS, "MASA" consortium, France.
What is the Cost and the Efficiency of Exposing Archaeological Data in the LOD?

There is a very great disparity in the organization of archaeological data and their management and archiving systems, when they exist. The immediate consequence of this state of affairs is the risk of the irreversible loss of a significant amount of inaccessible archaeological data. The MASA consortium has therefore set itself the objective of digitally sharing archaeological data by proposing guidelines to the archaeological community. There is an urgent need both to safeguard existing archival collections and ensure the re-use of old databases, and to ensure that emerging new databases use sustainable systems that will provide interoperability and the long-term reuse of data. The consortium's work therefore focuses on the classification and digitization of old archaeological archives, the documentation of its collections according to an appropriate structure, the re-use of old databases, the alignment of vocabularies used with standards, the matching of archaeological information systems with the domain ontology for cultural heritage (CIDOC-CRM, ISO 21127:2014), on the online publication of archives, data and syntheses linked with data.

The difficulties to be overcome are many, depending on whether structured databases or more informal batches are being processed, and must reconcile various situations such as:
- The digitization of old archives (often resulting from the work of only one archaeologist) which must be classified, safeguarded, made available and documented according to the archaeological value added. This involves producing an overlay to compensate for the absence of a data structure.
- The transformation of old databases developed with systems and in formats that are no longer accessible or disappearing, in order to preserve the data themselves and their structure when it exists.
- The interoperability of old, structured and still maintained systems, whose redeployment in standard and open formats is necessary but beyond the means available. While this may seem simpler, it is not the case and care must be taken not to delay those who have had advance notice.
- The development of new systems which must be designed directly in accordance with existing interoperability standards so that they do not have to be rethought in mid-term.

All the experiments carried out show the heuristic value of the operations necessary for the digital sharing of data, which is a reflective step in terms of both content and information structure. However, the final quality of the information shared according to these four processes varies. The creation of metadata on loosely structured corpora is a definite added value but never achieves the finesse of description of structured information systems. In addition, the use of metadata description standards alone does not offer the semantic enrichment that can be achieved by mapping with reference ontologies, which constitutes a production of knowledge in itself. In fine, exposing archaeological data in the LOD will help to build bridges with other heritage data but also with other themes.

These different processes and their consequences will be explored using a few examples.

3.3. Talk 3 by Fatiha Indman, PoitiersUniversity, “CAHIER” Consortium, France. CAHIER: "Bridges / Puentes" Between Text Sciences

CAHIER is a French consortium whose mission is to promote good digital practices in text sciences and to build a network of expertise in the SSH scientific community. The particularity of CAHIER is that it does not bring together research centers but projects. Project members aim at collectively finding or sharing solutions to digitize, edit, display and process their data. The purpose of the consortium is not to register data in a specific field but to create links between disciplines that use texts as their scientific objects and subjects: Literature, Linguistics, History, Philosophy, ICT, Computer science, etc.

CAHIER's approach propagates and disseminates practices that comply with the "FAIR principles", by acting upstream of the projects, in order to guarantee and promote the quality of the data and metadata that are produced.

Through the example of the "WebOai" metadata exhibit tool, developed by the Cahier consortium with the help of Huma-Num, we will show how the Cahier consortium prepares its digital corpus of sources (teiHeader) for dissemination, exhibition and research. WebOai implements an OAI-PMH repository from XML-TEI encoded data sources. We will show how the confrontation of methods and the exchange of solutions, within the consortium, have allowed researchers to reflect on their data quality and how we are contributing to building the "Digital humanities" community in France.

3.4. Talk 4 by Michel Jacobson, BNF-COCOon Platform-”CORLI” Consortium, France.
The Benefits of Data Linking and Use of the CoCOon Repository.

The first advantage of data linking is that it requires cleaning the data to make them sufficiently homogeneous for mass treatment, either automatic or assisted. In a repository where the applicant provides the description with little or no help and with little moderation, it leads quite quickly to alternative forms for the same resource. For instance, the identification of a person by name is not always normalized but is often subject to variants (case variant, order of elements, changes in civil status, use of a pseudonym, abbreviation, typing error, etc.). Moderation is sometimes made difficult also because of the use of foreign scripts or conventions.

Linking to a repository means that identification and description needs can be separated. For example, in a documentary resources repository, the actors involved in creating the document have to be identified in the document description. It will be advantageous to describe the actors in a distinct and specialized repository since the description templates for actors and documents won’t necessarily be the same. Moreover, an actor exists independently of his documentary production and other

repositories (of events, of objects, etc.) may have the same requirement for identification, making it interesting to share the service.

The criteria of coverage, governance and interoperability need to be taken into account when choosing a vocabulary. This choice is important because linking the data also means in some cases, deporting the task of description. In particular, the "collaborative" modes of governance of projects such as Geonames or Dbpedia have the advantage that one can directly enrich the repository to cover missing needs. One can also, as we have tried to do with the CoCOon (https://cocoon.huma-num.fr) platform dedicated to digital oral corpora, make producers or depositors responsible for enriching these vocabularies themselves. The choice of repositories rapidly proves to be strategic because they will be the linchpin in decompartmentalizing the data: either the repositories share common vocabularies, or their separate vocabularies are interlinked.

As part of the work on the CoCOon platform, for example, we have indexed a collection of speech recordings ("Speech Treasures") by aligning the themes present in their metadata with a Thesaurus (RAMEAU). This allowed us to 1) offer a new axis of navigation in the data, 2) bring these records closer to other cultural data (those of the BnF) which are indexed with the same thesaurus, 3) potentially bring together other cultural data through the alignment of RAMEAU with other reference systems such as the Dewey classification, the thesaurus of the German National Library, the National Library of Spain and the Library of Congress, 4) envision multilingualism with no added cost by exploiting this alignment between repositories, 5) facilitate the reuse of data and their discovery.

Normalized vocabularies are a bridge between repositories, making it possible to bring together isolated data and thus to give them a richer context, improving their readability. For example, the use of the Lexvo vocabulary - which includes all the codes of the ISO-639-3 standard - makes it possible to reconcile recordings, scientific documentation, geopolitical information, etc. for a given language.