"We": a Proposal for the TRIPLE platform

Stéphane Pouyllau (0000-0002-9619-1002)[1], Mélanie Bunel (0000-0002-7314-3892)[1], Laurent Capelli (0000-0002-1873-3857)[1], and Jean-Luc Minel (0000-0001-6253-6722)[2]

[1]TGIR Huma-Num UMS 3598, CNRS, Aix-Marseille Université, Campus Cordorcet, Paris, France
[2]Université Paris Nanterre, Laboratoire Modyco UMR 7114, Nanterre, France

September 18, 2020

Abstract

This research report is a proposal to define a platform, called We, for the TRIPLE project. First, the We platform is situated in the ecosystem of academic tools. Then its functionalities are described.

Important

This document proposes a conceptual representation of an innovative platform based on the discovery of experts, topics and projects using the analysis, classification, linking and enrichment of data. It is not a proposal of a Human-Machine Interface (HMI) which is under the responsibility of WP3 and WP5 as well as the editorialization of the components, but there are ideas and may constitute some areas of work to be discussed. Also, the different diagrams aim at illustrating the functionalities and not the design of the screens. These suggestions include elements of the proposal (experts), new public (Python library) and technological advances (Deep Learning) which were not so advanced when the proposal has been elaborated.

1 Introduction

The early 2000s saw the production of a very large number of studies on the analysis of researchers access practices to bibliographic information sources and documents (publications, serial data, etc.). In their study conducted in 2008 [21], the authors identified more than a hundred such studies. In their conclusion, the authors insisted on a point that seems still relevant today: *"A search engine is no longer the "place" where different cognitive structures are matched in interactions, but is seen as a more global system in which multiple variables come into play: the cognitive space of the actors, psychological, social and organizational contextual characteristics, and changing information needs. It is important to apprehend the user in an information-seeking situation in a much more global way than they are studied in cognitive models and, a fortiori, in the approach which still often underlies current studies on the use of search engines".*[1]

In the last ten years, after the arrival of Google Scholar, the number of current platforms has risen sharply. The study conducted by Know-Center, the TRIPLE WP7 leader [9], lists 47 platforms, with the arrival of private (ResearchGate, Academia) or non-governmental (Semantic Scholar[2], Dimensions[3]) stakeholders. Some of these projects are led by TRIPLE's partners (for example with Viper[4] developed by Open Knowledge Maps for OpenAIRE or Veille.science developed by Huma-Num with the search engine ISIDORE [26][5]). This study [9] corroborates the results published in 2016 by [13], namely that Google Scholar is the search engine used by 89% of users. Similarly, in the study carried out by a partner (Abertay University) of the TRIPLE project for deliverable D3.1 [8], Google Scholar is the search engine used by the majority of users interviewed. At the same time, usage studies have practically disappeared from the research work.

Without denying the importance of needs studies or usage studies [8], our proposals are rather based on feedback and targeted analysis of the ISIDORE platform's[6] uses as well as on the study of different platforms, such as Semantic Scholar or Dimensions. It should be noted that all these platforms claim to be based on technologies derived from Artificial Intelligence and more specifically from Deep Learning. Our proposal thus attempt to position a socio-technical device [1, 19] different from the existing platforms by specifying new functionalities in a complementary and non-competitive way.

General principle of the We platform ("**We**" for "**W**ho are the **e**xperts") is to boost the discovery of scientific issues in the fields of SSH by exploring connections between scholars (the experts), their scientific tracks (or evolutionary tracks of their work) and their productions (projects, publications, etc.). These tracks are the result of computations relying on Machine Learning techniques. They requires the commitment of the project partners to maintain efficiently and regularly updated semantic annotations. The We platform is part of development of "deep FAIR platforms"[29].

The We platform is positioned as the main access platform to the TRIPLE ecosystem. Accessible via a website, it provides access to information on the TRIPLE project, experts and information search tools in Social Sciences and Humanities (SSH) and value-added services developed by TRIPLE partners. Functioning as a "hub", it is in fact the single entry point for TRIPLE users.

[1] Translated from French by the authors of this report.

[2] `https://www.semanticscholar.org`

[3] `https://www.dimensions.ai`

[4] `https://openknowledgemaps.org/viper/`

[5] "Veille.science" is a proof of concept developed in 2016 around ISIDORE, see `https://doi.org/10.5281/zenodo.3965604`. Some elements are reused in the this proposal.

[6] `https://isidore.science`

This research report presents the lines of work of the Software Design and Information Management Team (GLIST) from Huma-Num[7] in terms of definition of functionalities of the We platform. The proposals are largely inspired from the work carried out for the design and implementation of the ISIDORE platform in France opened in 2010. In relation to the functional architecture [11], this report deals with the perimeter surrounded by the green rectangle in the figure 1.

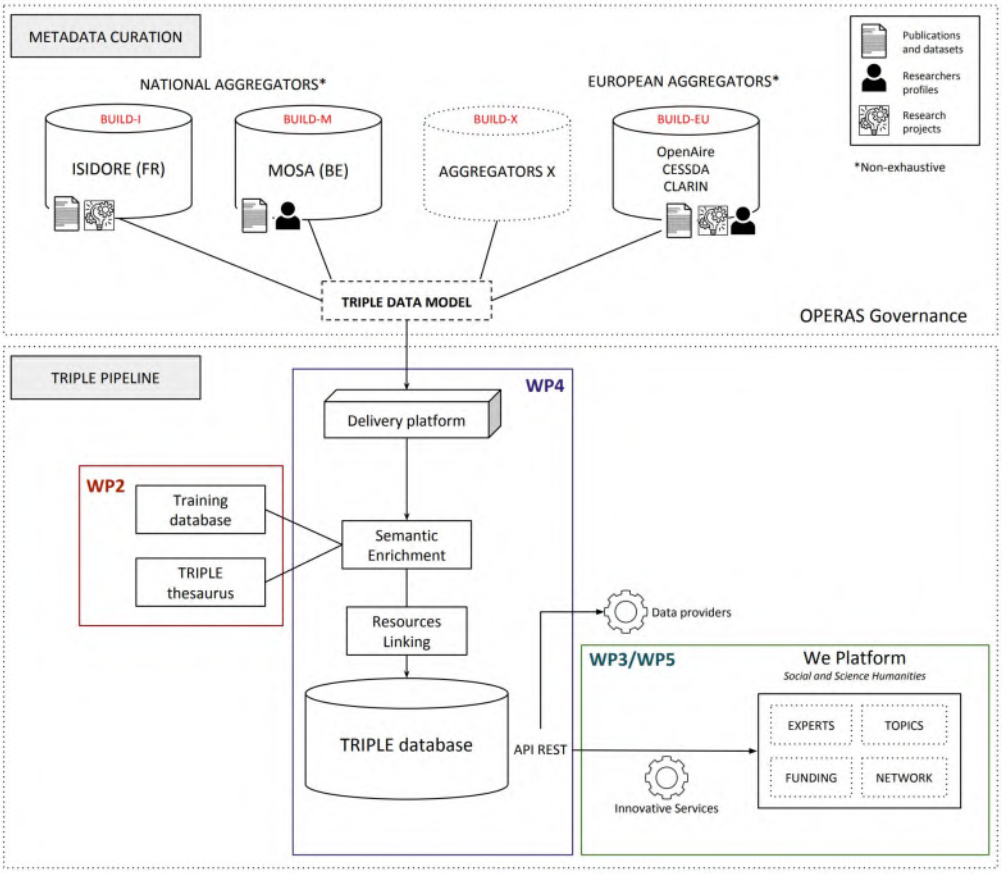

Figure 1: Functional architecture of TRIPLE and the We platform scope (diagram by M. Bunel. Source: Adapted from Deliverable 2.1: Data Acquisition Plan, 2020, TRIPLE project)

2 The Big Picture

2.1 Target Audience

In line with the TRIPLE proposal [27], the We platform is intended for :

- Scholars and PhD students from the SSH;
- Journalists (in order to diversify their sources, experts, etc.);
- Audiences seeking access to reliable information based on the notion of scientific community (include digital humanists who need advanced tools). For detailed audience, see the TRIPLE Deliverable: D8.3 Communication Strategy.[8]

It can be accessible via a sub domain URL of gotriple.eu (such as for example we.gotriple.eu).

[7] Huma-Num is a French Very Large Research Infrastructure with international reach devoted to Social Sciences and Humanities. See `https://www.huma-num.fr`

[8] `https://zenodo.org/communities/operaseu/`

2.2 Definition

The We platform is positioned as a socio-technical device [1, 2, 19] dedicated to the discovery of complementary scientific issues compared to generalist academic search engines (OpenAIRE, Google Scholar, Semantic Scholar, Dimension) but also more sectorial or disciplinary (such as ISIDORE for the SSH) or even compared to the discovery platforms provided to university libraries [6] by scientific publishers.

It exploits, via the TRIPLE API/REST, the enrichments and semantic links produced by applying Machine Learning methods in order to editorialize the contents of the TRIPLE database. In order to create semantic relationships between informational objects of different natures and coming from several European data aggregators, potential of data linking based on the principles of Linked Open Data (LOD) [5] are also exploited. As such, it does not take the form of a Web portal for documentary research, but offers discovery and navigation between scientific concepts ("Topics") and experts (see 4) in the sense of a set of research topics studied by a scholar. It should be noted that the linking is carried out upstream of TRIPLE's API/REST on the basis of a fine indexing of the metadata, of the full text when it is freely available.

In a nutshell, the We platform provides access to research data (publications, documents, skills, current or past projects, etc.) through an editorialization of the notion of experts and groups of them in SSH fields derived from TRIPLE's semantic processing[9]. In this way, We platform fully exploits the enrichment processing chain of TRIPLE[11].

2.3 Architecture

The We platform brings together on a single website all the services and tools giving access to enriched information as well as information relating to the life of the project. It contains (fig.2):

- The information flow on TRIPLE project news (for example via RSS feeds from a TRIPLE Web blog);
- A gateway, relying on SSHOC catalogue[10], giving access to descriptions of value-added services (via a database describing these services). To be included in this registry, a service must be generic enough and freely reusable for other platforms[11] ;
- Several web pages[12]enabling the discovery of scientific issues in the SSH fields. These pages contain all or part of TRIPLE's visualization, annotation, recommendation services, either in the form of an inclusion in We's Web UI, or as links pointing to a third party UI.

[9]Specific plateforms, like Association for Computing Machinery (`https://dl.acm.org/people`) propose interesting examples.

[10]`https://sshopencloud.eu/ssh-open-marketplace`

[11]It will depend on the policy of the TRIPLE project partners and on the positioning of these tools after the funded period of the project.

[12]Named w1, w2, ... in figure 3

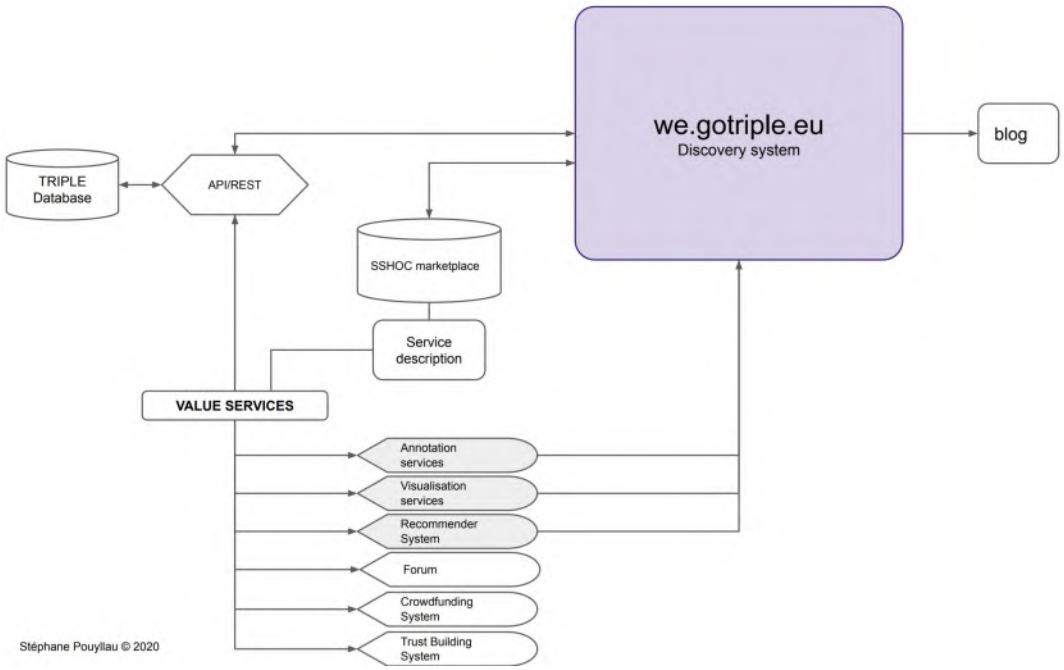

Figure 2: Simplified architecture of the We platform and its services (diagram by S. Pouyllau)

We detail below the functionalities offered by these web pages (fig.3).

- The ability to search, find and browse the work, activities of scientific experts;
- The ability to find projects and documents related to experts in one or more fields of research (see "Topics" below);
- The possibility to identify rising scientific emerging issues, innovative or early stage work, by using computed enrichment and networks of experts.

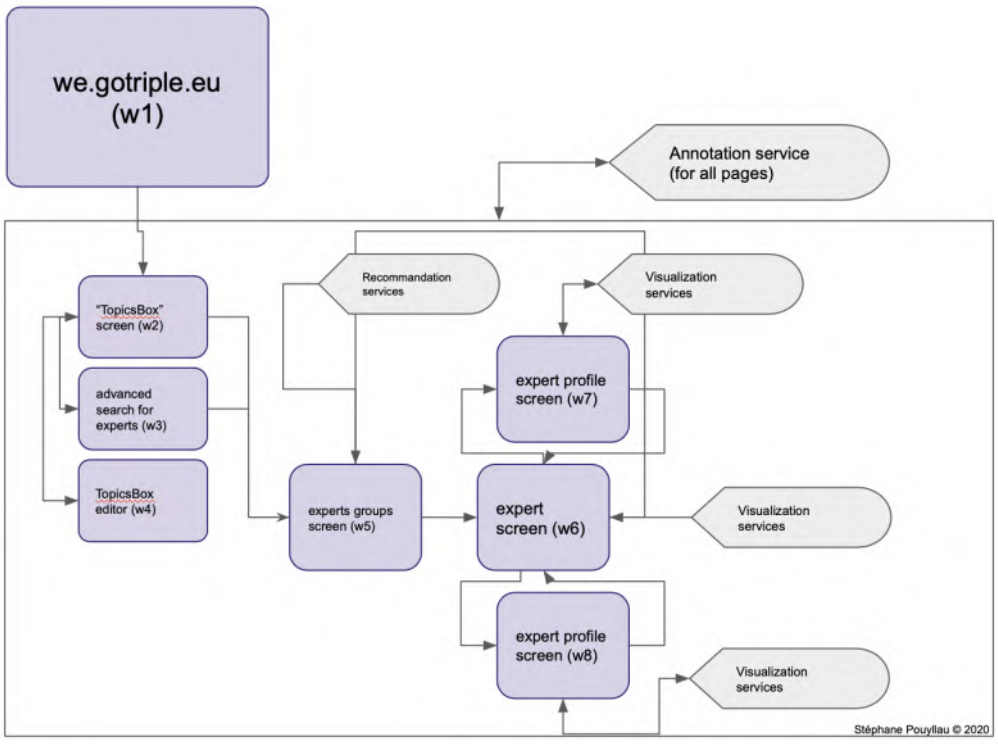

Figure 3: Detail of the functional pages of the We platform (diagram by S. Pouyllau)

The We platform will integrate different services such as visualization, recommendations, and annotations developed by the WP5 led by OKMaps/Nuromedia, and Net7, partners. Theses services will undoubtedly have to be accessible in the form of API results in order to be used in the Web interface designed by the WP4. These integrations will depend both of WP3 and WP5[13]. There are two modes of use of the WE platform, an unconnected mode (fig.4) and a connected mode (fig.5), with a specific web page for each mode.

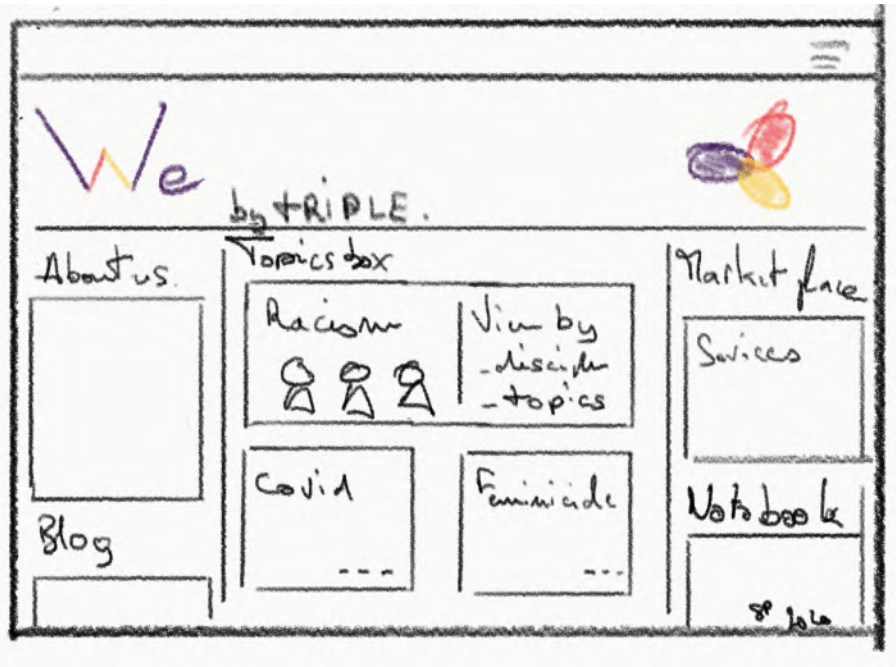

Figure 4: General picture for unconnected users (drawing by S. Pouyllau)

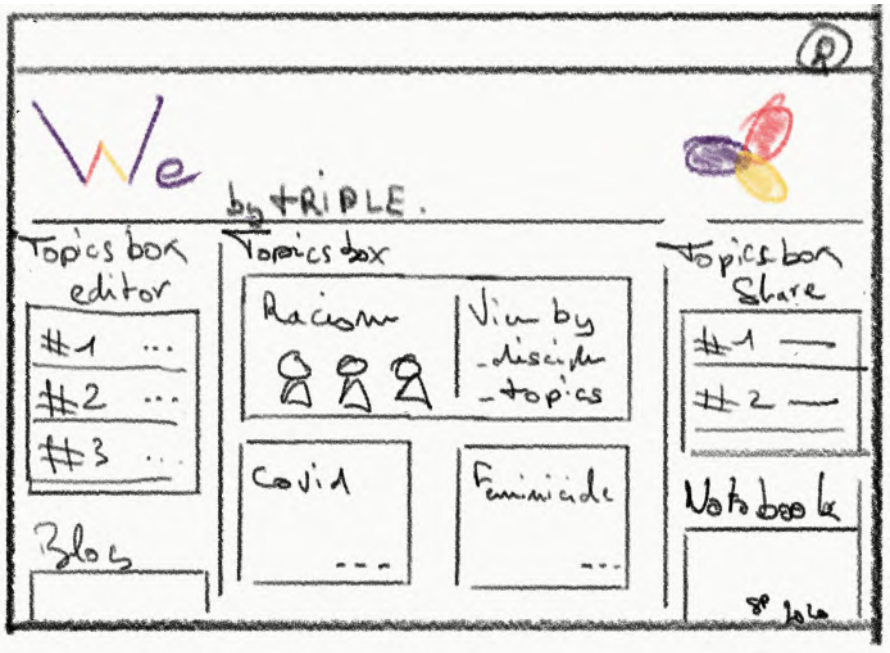

Figure 5: General picture for connected users (drawing by S. Pouyllau)

[13]It should be noted that at the time of writing this document, it is not possible to specify the modalities of integration in the Human-Machine Interface (HMI) We platform

3 Topics Box

"TopicsBox" (fig.6) are boxes in which a user specifies a scientific query for which he is looking for experts. The "TopicsBox" are the heart of access to TRIPLE's enriched information and the mandatory entry point to the rest of the discovery system. A scientific query combined concepts selected and edited by the partners of the TRIPLE project. Borrowed from scientific news or from the thesaurus[14] Library of Congress Subject Headings (LCSH) of the *Library of Congress* which are potentially related to several other concepts from repositories expressed in the LOD and aligned on Wikidata.

Figure 6: General picture of the concept of "TopicBox" (drawing by S. Pouyllau)

The editorialization of the TopixBox is under the responsibility of the WP3 with the support of the GLIST team.

3.1 Building TopicsBox in the Browser

TopicsBox are built by aggregating concepts searched in the TRIPLE database via the API/REST. An registered user can manage his TopicsBox in his library. We want to point out that TopicsBox combined topics (concepts) (fig.7) provided by the TRIPLE API, consequently an user can not express a query with his own keywords.

[14]The frequency of concepts updating will be decisive for the We plateform.

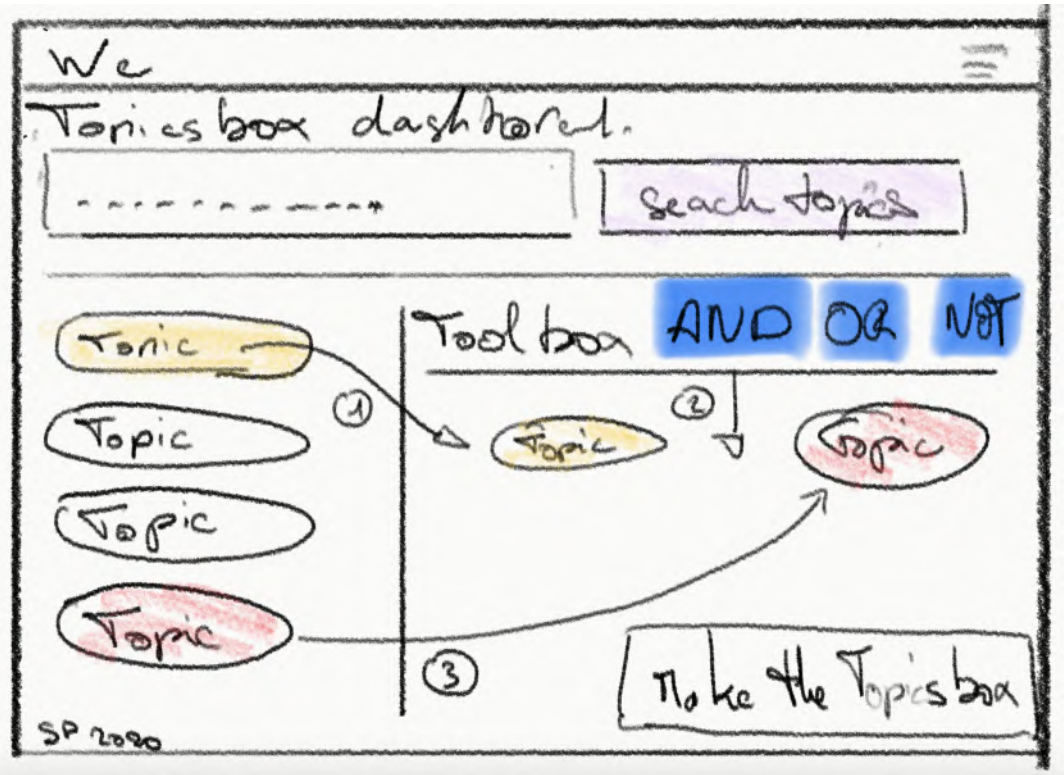

Figure 7: Editor of "TopicsBox" (drawing by S. Pouyllau)

4 Experts

4.1 Looking for an Expert

The notion of expert has given rise to an abundance of literature [12] as well as finding experts in academics [4, 7]. A survey on the state-of-the-art methods is presented in [20]. There are several reasons for using this type of system such as: recruiting reviewers, looking for scientific partners for research proposals, finding a PhD supervisor, and so on.

As described in [27, 8] looking for experts, relying mainly on university open archives, is a crucial task whose goals cover a wide range of needs in view of the target audience: scholars, students, journalists, and so on. As pointed out by [4] *"The use of university repositories is not so common, in spite of their wealth of bibliographic metadata of both local scientific records and student theses records. (...) One of the reasons for not exploiting institutional repositories on a greater scale could well be that they are often created, administered and run by librarians who generally are not research oriented and especially not towards computer science. Another reason might be that the data in the repositories often is entered by researchers and students themselves and therefore, in many cases, are not quality controlled."*

In the We platform, an expert is a highly active scientist who published scientific papers, produced scientific data sets or contributed to scientific projects. More specifically, a TRIPLE expert is defined by a certain volume of enrichments (disciplines, concepts) made from the semantic processing of the metadata of his publications, projects and description elements of his profile (description in his CV, etc.) and potentially from the processing of his scientific papers (depending on its availability).

A TRIPLE expert is characterized first of all by his/hers disciplines of expertise based on disciplinary categorization and annotation of his publications and/or projects. Then, his/hers links with other experts from the same disciplines and having one or more themes in common. These links may be upstream or downstream: "has been influenced by ...", "has influenced ..." or "is cited by ...". Finally, an expert belongs to a community (see

below the notion of group) having a sustained and/or regular co-publishing activity.

An expert can thus be characterized by a set of indicators that can be provided or computed by TRIPLE :

- An indicator of activity based on publication activity, membership of journal editorial boards ;
- An indicator of notoriety based on citations of the researcher's work.
- An indicator of participation in communities based on his/her participation in research projects.

The objective of this report is not to detail algorithms used to calculate a scholar's degree of expertise, neither how to rank them, but the prototype DIVA proposed by [4] could be a realistic source of inspiration. It should be noted that the computing of these indicators requires access to data not currently provided by aggregators (Isidore, OpenAire, Mosa, Narcis, etc.). Therefore, if these functionalities were retained, it would be necessary to engage discussions with these aggregators.

4.2 Displaying Experts

In the We platform, looking for experts is always conditioned by the expression in the query of topics and/or disciplines. Figure 8 shows an example of a sketch of screen dedicated to perform this kind of query and figures 9 and 10 propose some sketches showing results of this query.

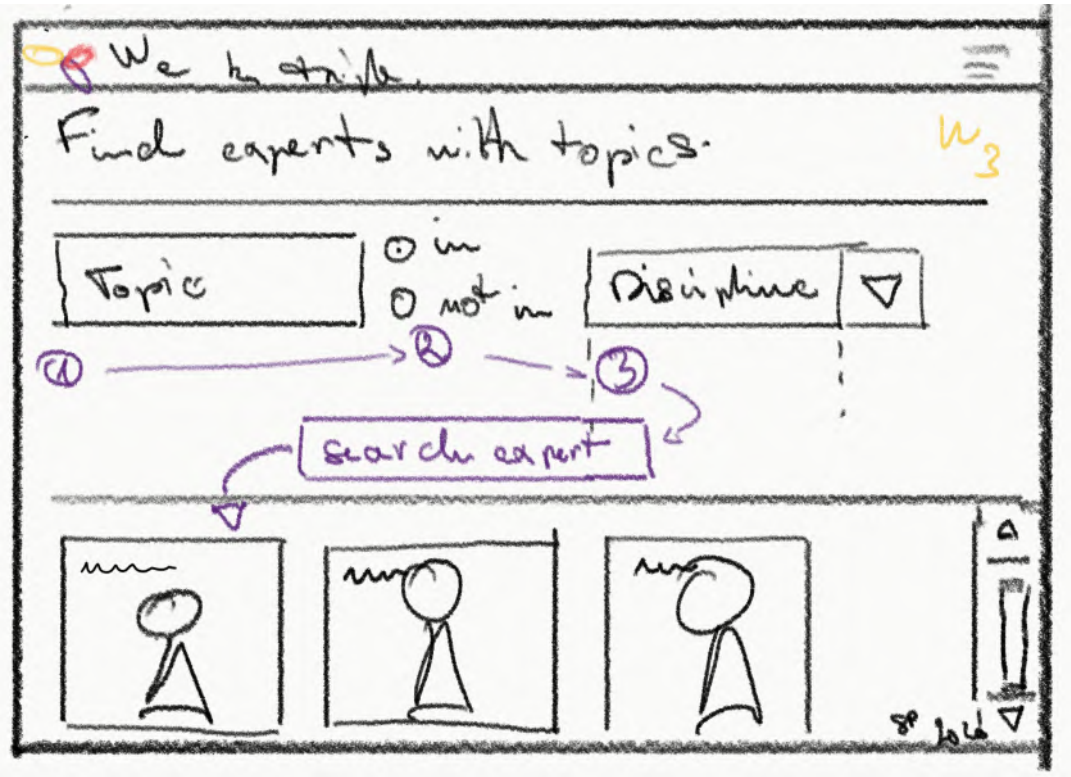

Figure 8: Looking for experts (sketch from S. Pouyllau)

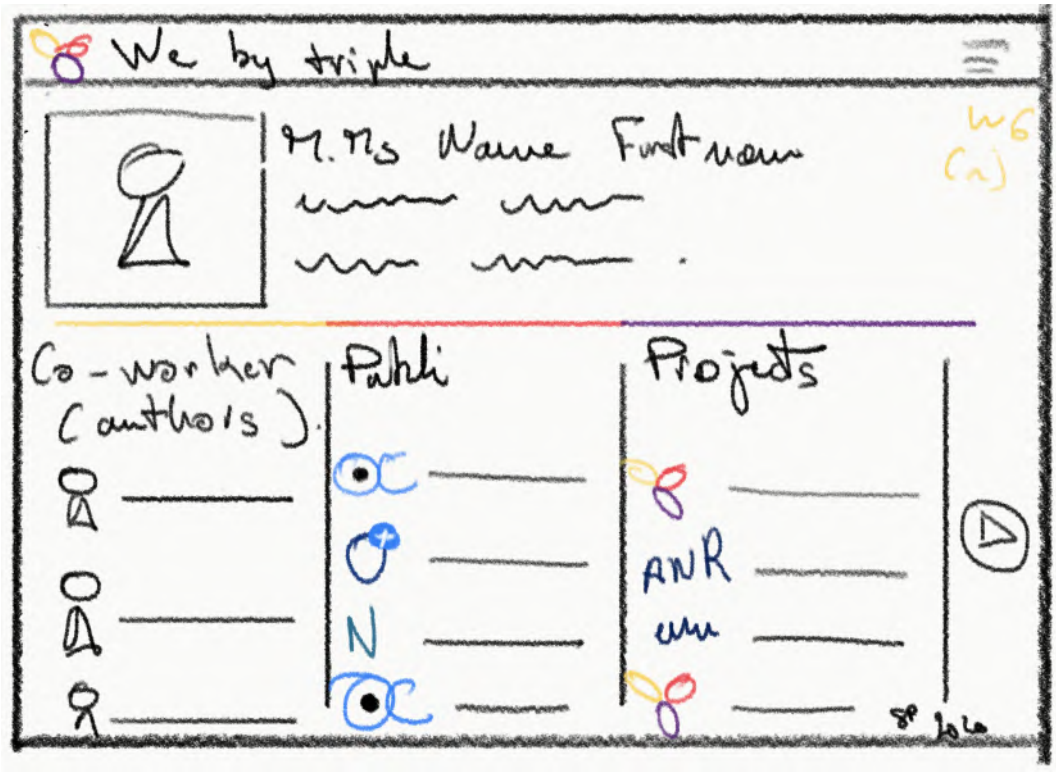

Figure 9: Finding experts (sketch from S. Pouyllau)

Figure 10: Finding experts (sketch from S. Pouyllau)

More precisely, the screen in figure 9 displays co-authors and publications of an expert, as well as projects in which the expert is or was involved. It should be noted that the links to publications are organized by aggregators (Isidore, OpenAIRE, Narcis, etc.) which allows the user to have a first appreciation of the expert's work. The screen in figure 10 displays projects in which the expert is or was involved and a list of scholars who influenced him. It should be noted that these data about experts could be enriched by

exploiting data not provided by TRIPLE. For example, these complementary data can be taken from different services mentioned by [15] such as Research Graph Foundation [15], Scholix [16] [28], a framework for Scholarly Link Exchange coordinated by a Research Data Alliance (RDA) Working Group, and the OpenAIRE Research Graph [22]. The FREYA project[17] mentioned in the TRIPLE WP6 led by Net7 partner [27] should become an important resource which should be used by TRIPLE.

5 Proposal of Innovative Concepts and Applications

5.1 Emerging Themes

As stated above, principles of the We platform is to boost the discovery of scientific issues in the fields of SSH through the links between scholars who are experts in topics, belong to well-known research communities (the groups) and their scientific tracks (or evolutionary track of their work) and their scientific productions (projects, publications, data sets, etc.). The notion of scientific tracks potentially leads to the notion of "emerging themes" bringing together several experts who do not necessarily have a link based on the co-signature of publications.

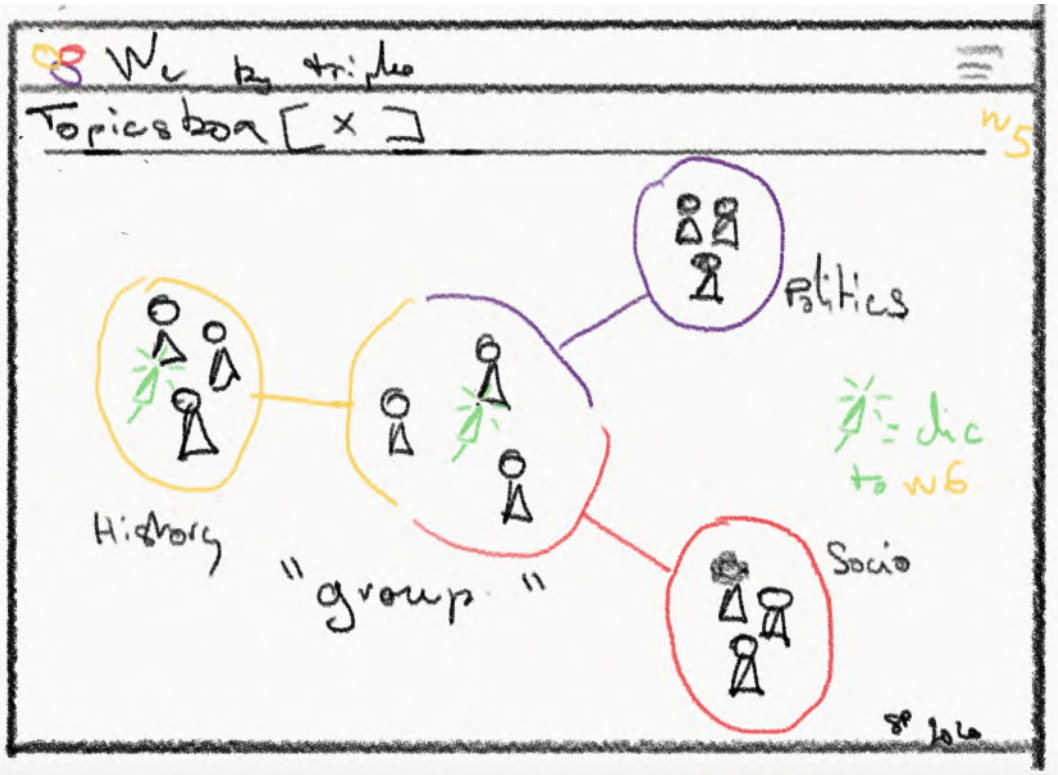

Figure 11: Expert groups and emerging themes (sketch from S. Pouyllau)

Belong to an "emerging theme" induces connections that are based solely on concepts, disciplines and any common enrichment between experts who are not co-authors. This notion is particularly interesting in the linguistic dimension of TRIPLE enrichment (9 languages). Discovering experts, who have never published together, from a pool of documents, projects or profiles, in 9 languages is one of the innovative and unprecedented features in the specialized field of SSH.

[15] https://researchgraph.org

[16] http://www.scholix.org

[17] FREYA is a 3-year project funded by the European Commission under the Horizon 2020 programme. The project aims to build the infrastructure for persistent identifiers as a core component of open science, in the EU and globally. FREYA will improve discovery, navigation, retrieval, and access of research resources https://www.project-freya.eu/en

Looking for emerging themes requires taking into account the evolution over time of the experts's research themes as well as the identification of groups to which they belong. There are several ways to identify experts on emerging themes. One approach would be to rely on Deep Learning techniques by exploiting open source machine learning framework such as Pytorch [18] or Keras [19]. More specifically, the classification of authors' titles and/or abstracts is an approach to be explored. In the same way as for the TopicsBoxes (see above), TRIPLE could offer the integration of a Jupyter Notebook for experienced users (detailed in Section 5.3).

5.2 Groups of Experts

Expert groups bring together experts from different disciplines who are not co-authors but work on related themes. An example of the representation of these groups and for an emerging theme is shown in Figure 11. We think that some propositions made by TRIPLE partner Open Knowledge Maps (OKMAPS) could be used and/or adapted to carry out this kind of searching. This suggestion needs further investigations which may be discussed later.

5.3 Building Embedded TopicsBox

With the idea of targeting digital humanists among the potential users of the platform, we propose to create a specific mode in order to editing "TopicsBox" in a programming environment in the form of Python libraries (or other) packaging the API/REST *notebooks* under the Jupyter environment with Python packages. Indeed, these environments are increasingly being used in the field of data science [17] and in the construction of data analysis applications, particularly in higher education [24]. This functionality will open up important possibilities for researchers mastering these programming environments.

For example, figures 12 and 13 show two Python scripts and their results that can be performed from the Triple SPARQL EndPoint via a Jupyter Notebook. The first example calculates the disciplines in which an author publishes and represents it with a pie chart. The second one computes the number of co-publications from an author[20]. Some plateforme use Python packages in Jupyter Notebooks, for example, DataCite, with GraphQL and PID Graph[16].

Our hypothesis is that a large number of scholars and PhD students master these techniques and that TRIPLE provides them a massive source of data on which it is possible to discover and process new research objects [18, 24, 14].

[18] https://pytorch.org

[19] https://keras.io

[20] Full examples are available and machine readable with Binder for Jupyter notebooks at `https://gitlab.huma-num.fr/spouyllau/isidore-jupyter/`

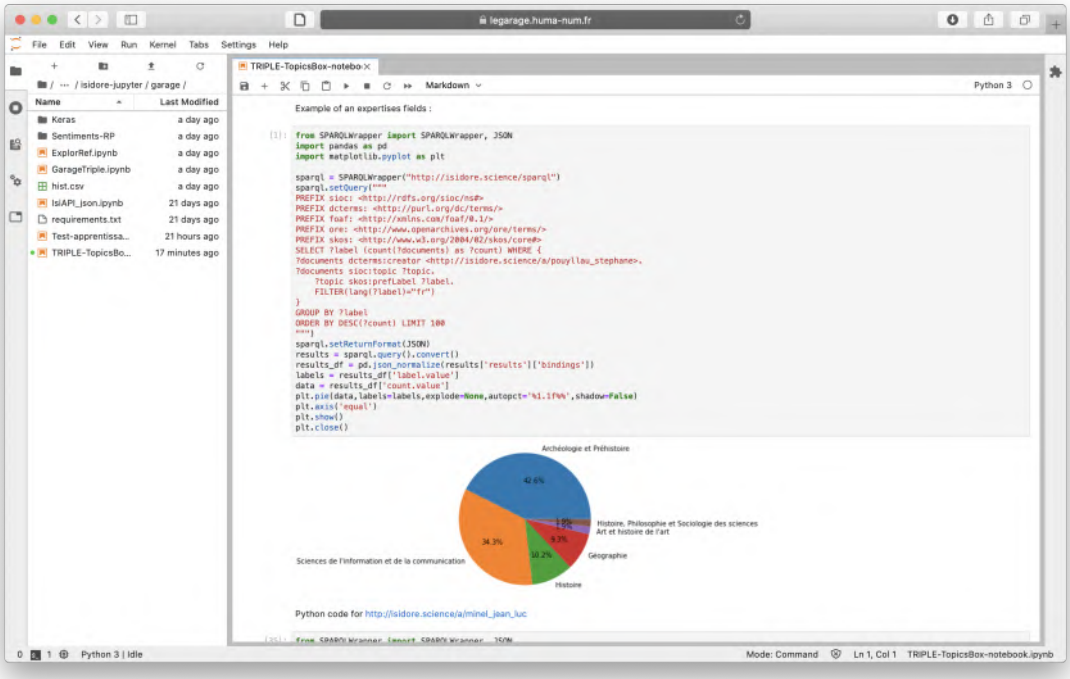

Figure 12: Computing an author's fields from the TRIPLE SPARQL EndPoint in a Jupyter notebook (screenshot from S. Pouyllau)

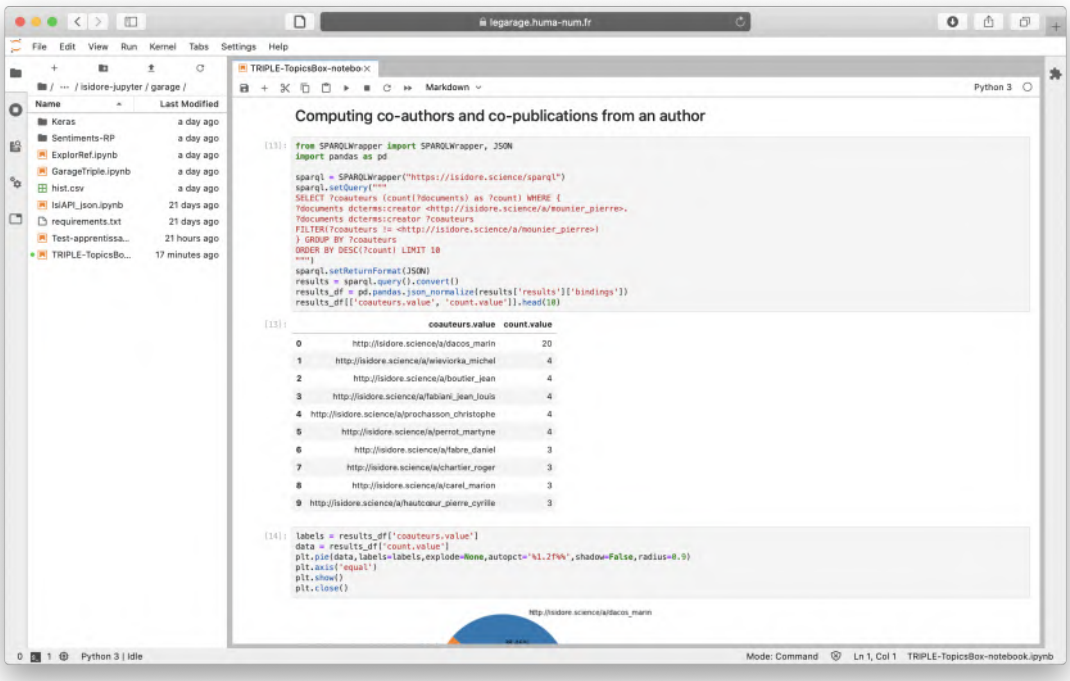

Figure 13: Computing co-publications from an author with the TRIPLE SPARQL End-Point in a Jupyter notebook (screenshot from S. Pouyllau)

6 Conclusion

We propose a socio-technical device, called We platform, to highlight and exploit the work of data collection, aggregation and enrichment carried out for many years by national search engines in the SSH disciplines or pan-continental portals who propose institutional sets.

This platform editorializes semantic enrichments carried out by the TRIPLE project by proposing the discovery of scientific issues in the fields of SSH by taking advantage of links between scholars (the experts), their known research communities (the groups) and their scientific tracks (or evolutionary track of their work) and their productions (projects, publications, etc.). This platform takes advantage of various services developed by TRIPLE project partners by including them either in a catalogue of services or directly in the HMI, i.e. in the User Interface Design (UI) to serve the User Experience Design (UX).

Contents

List of Figures

Glossary

GLIST Software Design and Information Management Team. 3

HMI Human-Machine Interface. 6, 14

LCSH Library of Congress Subject Headings. 7

LOD Linked Open Data. 4, 7

SSH Social Sciences and Humanities. 2, 11, 14

UI User Interface Design. 14

UX User Experience Design. 14

References

[1] Akrich, M.: Les formes de la médiation technique, Réseaux, 60, p. 87-98, (1993).

[2] Akrich, M.: The De-Scription of Technical Objects, in Bijker, W. & Law, J. Shaping Technology/Building Society. Studies in Sociotechnical Change, Cambridge MIT Press, 205-224, (1992).

[3] Aryani, A., Poblet, M., Unsworth, K., Wang, J., Evans, B., Devaraju, A., Hausstein, B. Klas, C-P., Zapilko, B. Kaplun, S.: A Research Graph dataset for connecting research data repositories using RD-Switchboard, Scientific Data, 5 (2018). 10.1038/sdata.2018.99

[4] Angelova, M., Devagiri, V., Boeva, V., Linde, P., Lavesson, N.: An expertise Recomender System Based on Data from Institutional Repository (DiVA), ELPUB, Canada, (2018).

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

[6] Bourdenet, P.: L'espace documentaire en restructuration : l'évolution des services des bibliothèques universitaires, (2013). 10670/1.lnieuv.

[7] Cifariello, P., Ferragina, P., Ponza, M.: Wiser: A semantic approach for expert finding in academia based on entity linking. Information System. 82, 1–16, (2019).

[8] Forbes, S, De Paoli, S., Błaszczyńska, M., Maciej M.: Deliverable D3.1, Report on user needs. TRIPLE Project, (2020). 10.5281/zenodo.3925022.

[9] Breitfuss, G., Barreiros, C., Demouchel, S., Blotiere, E., Bouillard, M., Di Donato, F., Forbes, P.: Deliverable D7.1, Report on Stakeholder and Opportunity Analysis TRIPLE Project, (2020). 10.5281/zenodo.3925662.

[10] Bunel, M., Capelli, L., Minel, J.L, Pouyllau, S.: An ontology for the TRIPLE Data Model, (2020). gitlab.huma-num.fr/triple/model.

[11] Bunel, M., Capelli, L., Minel, J.L, Pouyllau, S.: Thinking about the Architecture of the Discovery Platform of the TRIPLE Project, (2020). halshs-02889863.

[12] Bessy, C., Chateauraynaud, E.: Experts et faussaires. Pour une sociologie de la perception, Editions Pétra, coll. "Pragmatismes", 2e édition, 522 p., ISBN : 978-2-84743-088-2 (2014).

[13] Delgado Lopez-Cozar, E., Orduna-malea, E., Martin-Martin, A.: Google Scholar as a data source for research assessment, Computer Science, (2018). 10.31235/osf.io/pqr53.

[14] Edmond, J. (ed.): Digital Technology and the Practices of Humanities Research, OpenBook Publishers, (2020).

[15] Fenner, M., Aryani,A.: Introducing the PID Graph, (2020). 10.5438/jwvf-8a66

[16] Fenner, M.: Using Jupyter Notebooks with GraphQL and the PID Graph, (2019). 10.5438/hwaw-xe52

[17] Herta, C., et al.: Deep Teaching: Materials for Teaching Machine and Deep Learning. Repositorio Institucional de la Universitat Politècnica de València, (2019). 10.4995/HEAD19.2019.9177

[18] Jacob, C.: Pour une herméneutique numérique en sciences historiques, (2019). lieuxdesavoir.hypotheses.org/1549

[19] Latour, B.: La Clé de Berlin. Petites leçons en sociologie des sciences [1993]. Paris : Seuil (Points sciences), (1996).

[20] Lin, S., Hong, W., Wang, D., Li, T.: A Survey on Expert Finding Techniques, Journal of Intelligent Information Systems. 49(2): 255-275, (2017). 10.1007/s10844-016-0440-5

[21] Ihadjadene, M. et Chaudiron, S.: Quelles analyses de l'usage des moteurs de recherche, Questions de communication, 14, (2008). 10.4000/questionsdecommunication.604.

[22] Manghi, P., Bardi, A.: The OpenAIRE Research Graph - Opportunities and challenges for science, International Open Science conference, Berlin (2019). record/2600275#.XzFV437gqCg

[23] Pouyllau, S., Minel, J.L, Kilouchi, S., Capelli, L.: Bilan 2011 de la plateforme ISIDORE et perspectives 2012-2015, (2012). 10670/1.bqexsj.

[24] Pouyllau, S.: L'exploitation d'ISIDORE et de son SPARQL endpoint avec Jupyter, (2019). 10670/1.cztpk9

[25] Pouyllau, S.: Les moteurs de recherche profitent aussi de la sémantique, (2012). 10.3917/docsi.484.0022 également disponible en libre accès via 10670/1.oodgbq.

[26] Pouyllau, S.: Veille.science : une veille et intelligence scientifique en SHS à destination des décideurs, (2016). 10.5281/zenodo.3965604

[27] Transforming Research through Innovative Practices for Linked interdisciplinary Exploration TRIPLE. The European discovery platform dedicated to SSH resources, proposal 863420, (2018).

[28] Burton, A., Koers, H., Manghi, P., Stocker, M., Fenner, M., Aryani, A., Bruzzo, S., Diepenbroek, M., Schindler, U.: The Scholix Framework for Interoperability in Data-Literature Information Exchange, D-Lib Magazine, Vol 23(1/2), (2017).

[29] Parland-von Essen, J., Riungu-Kalliosaari, L.: 2nd Report on FAIR requirements for persistence and interoperability (D2.4), 2020, FAIRSFAIR project. 10.5281/zenodo.3557381