# FAIRifying a scholarly publishing service:  Methodology based on the OpenEdition's internal FAIR audit

## Introduction

The FAIR principles—Findability, Accessibility, Interoperability, and Reusability—are guidelines whose aim is to improve the management of digital scholarly resources for both humans and machines. The principles define the characteristics that enable discovery and reuse of data and, more broadly, any type of digital research object (tools, algorithms, workflows, etc.). They assist different research actors (such as researchers, data stewards, service providers) to assess and increase the degree of FAIRness of their data. The barriers to the FAIR principles' implementation remain low: the principles are concise, domain-independent, and high-level. The constitutive elements are related, yet separable, and they can be combined in different ways.

Initiatives for the adoption of FAIR principles have predominantly targeted data producers, including researchers and data stewards. However, it is also widely acknowledged that the services providing the data should themselves be FAIR-compliant. As the FAIRsFAIR report (Koers et al., 2020a) on the FAIRness of services stated, "data and other digital objects cannot be made FAIR without several enabling services that facilitate the provisioning of persistent identifiers (PIDs), provide indexable resources and support access, amongst other factors". Although there are existing and valid frameworks to assess the FAIRness of data repositories, the FAIRsFAIR report also noted that "for data services other than data repositories the current landscape is less populated".

One significant example of service that has been working to comply with the FAIR principles but that literature commonly neglects, is the publishing service. Although usually considered as the conclusive part of research, scholarly publishing and communication is actually at the heart of scientific activity. The principles designed for the improvement of management and circulation of research data should therefore apply also, with the appropriate adjustments, to publishing data and services. This is particularly true in the context of Social Sciences and Humanities (SSH), where the main research output is often a publication, and textual corpora constitute in many cases the primary data of the research projects. In that prospect, even non-traditional publications, such as blogs, online annotations, or scientific events announcements, represent data that can undergo a FAIRification process and potential research objects. The full integration of scholarly publishing within the research lifecycle is confirmed by the recommendations coming from the Plan S, at a European level, or the

National Plan for Open Science (Plan National pour la Science Ouverte), in the French context, both inspired, directly or indirectly, by the FAIR principles.

We can list various motivations for the FAIRification of a publishing service. It relates to the development of the Open Science environment briefly described above, in which publications can be considered as data. Therefore, publications can integrate this environment through the application of FAIR principles. Furthermore, the FAIR principles make it possible to address some of the challenges encountered by the publishing services in terms of data management, metadata generation, and interoperability. More specifically, publishing services providing open access contents can find in the FAIR principles a useful tool to technically support their objective of openness.

All these aspects led OpenEdition[1], an organization maintaining four different publishing platforms, to conduct a FAIR internal audit in 2019. OpenEdition is a French digital infrastructure for open scholarly communication in the SSH domain that brings together four complementary platforms focused on journals (OpenEdition Journals[2]), book series (OpenEdition Books[3]), research blogs (Hypotheses[4]), and academic events (Calenda[5]). This paper stems from the FAIRification work conducted by the OpenEdition team and presents the lessons learned. It provides an example of FAIRification of a publishing platform to other comparable services. Without entering in all the details of the FAIR assessment, this paper describes the main components of the methodology used by OpenEdition during its FAIR internal audit. It was the birth of an on-the-job yet reasoned and efficient methodology. From the overall methodology used by OpenEdition, which was based on an adaptation of the FAIR principles to the publishing specific context, we believe that a generic framework can be extracted and reused by other publishing services. It is planned indeed, in the context of the OPERAS[6] European Research Infrastructure coordinated by OpenEdition, to create a toolkit for the FAIRification of publishing services, that we will present briefly in conclusion.

## 1. Why and How to FAIRify a publishing service?

### 1.1 Scholarly publishing and FAIR

Electronic scholarly publishing is well acquainted with the main aspects of data provision and management, thanks to its objective of dissemination and its rather broad use of persistent identifiers and metadata standards. However, the level of technical readiness of

[1] https://www.openedition.org/

[2] https://journals.openedition.org/

[3] https://books.openedition.org/

[4] https://fr.hypotheses.org/

[5] https://calenda.org/

[6] https://www.operas-eu.org/

the overall publishing landscape remains uneven and the evolution of publishing formats and objects creates new challenges for data management. Furthermore, the increasing commonalities between publishing systems and data repositories and between publications and datasets require to be directly addressed in order to facilitate their smooth convergence. In this context, publishing services can be seen as the datafication unit of the publications and the FAIR principles appear to offer an appropriate tool to consistently integrate publications into the digital research environment. Indeed, the generic principles of Findability, Accessibility, Interoperability, and Reusability could guide, and in many cases they already do, publishing systems. These four principles provide an analytical grid that is applicable also to this kind of services. This concerns primarily the metadata (identifiers, bibliographical information, controlled vocabularies), but also the content, here intended as the data (content's accessibility, standard formats, licensing). In that sense, the FAIR principles allow to obtain a global overview of publishing service provision in terms of technical characteristics and quality.

However, the FAIR principles were firstly designed for data, more specifically for research data, which face specific management challenges. The FAIR principles aim at facilitating the exchange and the combination of wide and complex research data, either by humans or by machines, thus with a focus on machine-readability. Although the FAIR principles correspond to publishing activities at a generic level, they necessitate some adjustments when considering publications. Publications, as expressed by various legal texts, are a product "of the mind", which means that humans and human relationships still have a major role in their exchange, distinct from machine-readability purposes. Reusability, for instance, is by definition always ensured for publications: reading texts, which is the goal of publishing, is a rather satisfactory form of content reusability that does not require digital system repeated updates. In the same way, Accessibility, in the sense of making contents available, openly or not, is of course at the heart of digital publishing, while often only part of research data is made available. This focus on human activities and relationships also imply, however, that the datafication aspect of publishing services can be undermined. In fact, publishing services provide "content" more than they provide "data": a PDF file, which is sufficient for human readers, is digital data in a minimal sense, as it is not an interoperable format easily processable by machines. In the same way, the persistent identification of published contents often concerns only the final output. The management of versions and provenance information of a publication is therefore different than the one of a dataset constituted, for instance, both of primary and secondary data. Indeed, from the point of view of the service, the dynamics and workflows of a publishing service are very specific and distinct from those of research data creation. For all these reasons, the FAIRification of a publishing service requires some adjustments, first in terms of methodology.

## 1.2 A general methodology for FAIR publishing services

In this paper, we understand methodology as a broad concept, covering the various steps taken by a publishing service for its FAIRification. The landscape of FAIRification and FAIR self-assessment tools is now widely populated, but such tools address mostly the research datasets or data repositories. The evaluation that they provide does not fully apply to publishing contents or to the specificities of publishing services. As reported by the FAIRsFAIR report, the many existing FAIR-scoring tools would have proved insufficient to accurately "consider [the] several dimensions of a service, i.e., not only functional aspects ('utility' in FitSM terms) but also aspects that speak to quality, documentation, sustainability" (Koers et al., 2020a). In the same way, certification frameworks for data repositories, such as CoreTrustSeal[7], which would prove efficient FAIRification tools in a further stage, are hardly applicable to publishing services in a transitional process towards datafication.

The methodology presented here is based on the work conducted within OpenEdition. The paper does not intend to fully report on this work, but rather take a step back and consider this whole process as an empirical method which could be formalized and then reused. The methodology, therefore, does not concern only the FAIR assessment of OpenEdition, but all the main phases of the FAIRification process. For readability purposes, the phases are not presented as they chronologically happened, but their content remains unchanged.

Some components of this methodology are common to many other assessment projects, for instance, a study of the context, the analysis of some actual use cases, and the prioritization of the tasks. Two aspects, however, characterize this methodology: the specific combination and articulation of the various activities carried out at a publishing system (in this specific case, OpenEdition) and the use of the FAIR principles as an analytical grid.

The first phase, preparation, gathered the available information able to define the perimeter of the FAIR review. It consisted of an analysis of both the external and the internal contexts. The external context reveals the scholarly publishing landscape. Its study allows a publishing service to understand the aspects related to the open science environment, the FAIR principles themselves, and the initiatives specific to the scholarly publishing environment, like Plan S. This study also comprised defining some relevant publishing-related concepts, such as persistent identifiers and licenses. All these elements are already part of the everyday life of a publishing system; however, it is important to verify if all the aspects are well understood, how they relate to the activities carried out, what is already in place and what should be improved, among other things. The external context is presented at the section Landscape study and definitions

[7] https://www.coretrustseal.org/

The internal context sheds light on the service provision. At this step of the analysis a few actual use cases were mapped (the service and the use cases are detailed in the section OpenEdition's context). These use cases correspond to some situations experienced by OpenEdition, but it could be the case of any publishing service, and its analysis allows to draw a list of potential service improvements that could be achieved by implementing FAIR principles.

The second step was the assessment phase, the most extensive one. It comprised distinct steps. The first one consisted in contextualizing the FAIR principles, in other words: applying the generic FAIR principles to the context of open access scholarly publishing at a rather general level. The second step listed the distinct datasets that the review would consider. The FAIR analytical full review constituted the third step in which each dataset was analyzed thoroughly according to the 15 detailed recommendations[8] that derive from the four foundational FAIR principles (Wilkinson et al., 2016). As we can see, the general process of the review progressively increased the level of precision. Where the analysis revealed that more specific information was lacking to ensure a complete FAIR implementation, specific synthetical assessments were conducted. These assessments relied on a technical state of the art and a contextual analysis of the current status in the organization.

Based on the content of both the preparation and the assessment phases, the last phase consisted in producing a list of recommendations for the FAIR principles' implementation. Such recommendations imply other actors than the authors of the review, namely the other members of the organization and its customers. The recommendations comprised a plan of actions and further steps to be envisioned. The plan of action relied on a selection of the areas where FAIRification could be improved, whereas the further steps represented the classification of the objectives according to the service priorities in terms of feasibility, utility, and warranty.

## 2. OpenEdition's context

### 2.1 OpenEdition's platforms[9]

OpenEdition provides publishing services to publishers and authors for four types of contents. The organization does not take charge of the editorial process. It relies on the conversion of textual files into the TEI format and on a built-in Content Management System (Lodel), allowing for the creation of rich metadata. The organization ensures the display of

[8] The four foundational FAIR principles are further specified through 15 recommendations, available here: https://www.go-fair.org/fair-principles/.

[9] https://www.openedition.org/10918

the publications on its platforms and the dissemination of the metadata in various indexing services.

The Journals platform is dedicated to scientific journals in the humanities and social sciences and it gathers more than 500 publications. It promotes academic electronic publishing and open access. Journals may apply to join the initiative. Although journals have to meet some requirements to be admissible, they may decide on having an electronic-only format or also keeping a printed version, they may maintain their financial positions and they all have their own peer review committee.

The Books platform aims at building an international library for the digital humanities, and encourages publishers to develop Open Access in the long-term. It offers 10,000 books, of which three quarters are in Open Access, from more than 100 different publishers.

The scientific blogs platform, Hypotheses, hosts more than 3500 blogs of various types: research, fieldwork, seminars, etc. All its content is in Open Access. Hypothese uses the free and open-source content management system, Wordpress software.

Finally, Calenda is an online platform dedicated to research news, prioritising conferences, seminars, calls to contribution, research grants offers, etc. It has published more than 45,000 events in Open Access.

The infrastructure uploaded and published more than 900,000 documents in the year of 2020, most of these documents are in open access. Table 1 displays some numbers reporting OpenEdition's results in the last two years.

Table 1: Number of documents uploaded and published in OpenEdition's platforms in 2019 and 2020

| **Documents uploaded and published** | **2019** | **2020** |
|---|---|---|
| books | 9,000 | 10,000 |
| journals | 500 | 550 |
| research blogs | 3,200 | 3,700 |
| scientific events | 43,000 | 45,000 |
| total of documents | + 8000,000 | + 900,000 |

| | | |
|---|---|---|
| % in open access | 95 % | 96 % |

Source: OpenEdition reports, available at https://www.openedition.org/25480

## 2.2 FAIR-related use cases from OpenEdition

A careful observation of OpenEdition daily activities, allowed the team to list a series of actual use cases that took place in the infrastructure's everyday life. These use cases illustrate situations where a systematic application of the FAIR principles would have been beneficial and reveal how the infrastructure could gain by implementing them or lose by not doing so. The cases regard primarily the practices related to identifiers and licenses.

The first use case regards the existence of parallel identifying systems, with DOIs on one hand, and internal identifiers based on the OAI-PMH[10] repository, on the other hand. Documentary units are identified internally with reference to the platform (for example, "journals.openedition.org/archeomed/7020" and "oai:revues.org:archeomed/7020", respectively). Therefore, if the name of a platform and hence the URLs should be modified it would lead to an identifier modification, contrary to the principle that identifiers should be persistent. This was the case for the platform Journals that used to be called Revues. Similarly, modifying the name of a journal and hence the corresponding URL would probably also be easier to resolve with a persistent identifier.

A second example of the benefits of applying FAIR principles is the case of unpublishing or removing records. When it happens, the content's record is deindexed from the system's database that is used to feed the OAI-PMH repository. The content is no longer available in the OAI, but the information on deletion is not recorded. As a result, the resource remains listed in the referencing services that harvest the OpenEdition's OAI repositories (such as Isidore[11]) and point to URLs that no longer exist, giving a 404 response. Similarly, when deleting a document, the DOI resolution cannot point to metadata nor indicate that the resource has been deleted.

Another use case concerns the type of reuse license that is applicable to the content. In cases of reuse requests, the organization has generally been incapable of providing a clear answer to an applicant on the type of reuse they are entitled to make of the contents. This concerns in particular the full text TEI version of the content. The application and clear display of a user license (FAIR R1.1) would rectify this problem. For illustrative purposes,

[10] Open Archive Initiative - Protocol for Metadata Harvesting (OAI-PMH) is an open protocol for harvesting of standardized metadata. It relies on a repository where harvesters collect metadata.

[11] Isidore is a French search engine dedicated to the SSH. It is maintained by the Research Infrastructure Huma-Num, a close partner of OpenEdition. See: https://isidore.science/.

we could cite some situations that could benefit from an explicit license: access to the full text for indexing purposes; access to the full text for republication purposes; PDFs version republication; republication of an annotated corpus based on OpenEdition's contents.

The last use case refers to the identification of the publications' authors. OpenEdition was asked to provide the record of the publications produced by professors and researchers from a specific university. Even with the list of authors (surname, first name, structure), OpenEdition's system was only able to provide an unreliable list of publications. Better identification of the authors (FAIR I1) would undoubtedly have made it possible to respond more reliably to this request.

We could say that, at that moment, OpenEdition did not have all the required information to address the use cases. Nevertheless, looking for the answer helped to specify the FAIRification priorities.

## 3. Landscape study and definitions

### 3.1 The Open Science environment

A clear appreciation of the FAIR assessment of a publishing system depends not only on the awareness of the Open Science context and its related concepts, but also on the understanding of how such concepts relate to FAIR principles.

As OpenEdition develops open access digital publishing, it is crucial to understand the relationship between FAIR and openness. FAIR is clearly distinct from open in order to ensure the security of sensitive data or protected resources, and it presents itself as a technical common ground enabling various dissemination policies. However, not only the FAIR principles are often used in connection with open science, especially in our context of open access publishing, but they also share some requirements with recommendations that are distinctive of the open science movement. The landscape study precisely helped to assess the convergences and differences between FAIR and openness.

Open science is a growing movement to make scientific processes more transparent and publications and data more available. Put differently, it aims to build a whole ecosystem in which science will be more cumulative, more supported by data, and able to provide universal access to the produced knowledge. The notion of open science turns around a few concepts, such as open data, open access, open methodology, and open source.

Investigating this landscape, with the FAIR principles as a starting point, we identified a few notions that share comparable and sometimes identical recommendations. All together they compose, for various stakeholders and policy makers, the open science environment where scholarly publishing services also take place.

*FAIR principles*

One of the ways to further enhance open science practices is by structuring research data and publications so that they can be found, accessed, and reused. The FAIR Principles formulation helped to further this movement by specifying the minimum requirements for research products to be reusable, verifiable, and citable. The FAIR principles "emphasise machine-actionability"[12] and are founded on the idea that it is the ability to connect information that gives it meaning and enables its reuse. Since their first appearance, the principles have become an integral part of the various definitions of open science.

At an international level, the FAIR principles implementation is supported by GOFAIR and the Research Data Alliance (RDA)[13]. At a European level, the construction of the European Open Science Cloud (EOSC)[14] strongly relies on FAIR. With the French National Plan for Open Science (Plan National pour la Science Ouverte—PNSO)[15], renewed and reinforced in 2021[16], France has adopted an ambitious policy committed to making research results open to all. To meet this end, three axes have been conceived, being one of them explicitly related to the FAIR principles: "ensure that data produced by government-funded research in France are gradually structured to comply with the FAIR Data Principles". More generally, the PNSO stresses the importance of integrating the national development of open science with the international actions of the aforementioned EOSC, GOFAIR, and RDA.

*Open data*

The framework of the FAIR principles relates to another concept: open data. The notion of open data is connected to the notion of knowledge. Knowledge is only open if anyone can freely use it, reuse it, modify it, and share it. A few principles, presented on the Open Data Handbook[17] constitute the basis of open data. The Handbook focuses on three main axes: Availability and access, Re-use and redistribution, and Universal participation. The second one, outlining the need for licenses that allows re-use, redistribution, and link with other data, is close to the FAIR principles. Regarding the access to the resources, the FAIR principles do not recommend openness, but accessibility, i.e., the technical possibility to access the resources in a consistent and robust way, even under conditions. For this very reason, however, the FAIR principles implementation can also support the development of open data.

[12] GOFAIR, "FAIR principles": https://www.go-fair.org/fair-principles/.

[13] https://www.rd-alliance.org/

[14] https://eosc-portal.eu/

[15] https://www.ouvrirlascience.fr/national-plan-for-open-science-4th-july-2018/

[16] https://www.cnrs.fr/en/node/5883

[17] https://opendatahandbook.org/guide/en/what-is-open-data/

### *Plan S*

Another element to be considered is Plan S[18], which has a specific status in our scenario. Plan S was established by a consortium of funders and research organizations and, since 2021, it has mandatory value for the journals funded by the members of the consortium. The plan is structured around ten principles, with additional guidance regarding technical requirements. Convergences with the FAIR principles appear clearly in some Plan S principles, especially in the first point of Plan S, concerning the use of open licenses such as Creative Commons (CC) and the FAIR principle "(Meta)data are released with a clear and accessible data usage license". Among the technical criteria that are mandatory or recommended by Plan S there are other concerns shared with FAIR:

- use of a persistent identifier (FAIR F.1);
- present metadata related to sponsors (FAIR F.2, R.1.2);
- metadata should be under license CC0 (FAIR R.1.1);
- utilise a machine-readable format as JATS, TEI, etc. (FAIR R.1.3).

### *Linked Open Data*

The last fundamental notion related to the FAIR principles is a technical one, Linked Open Data (LOD)[19]. LOD is a set of design principles for sharing machine-readable interlinked open data. According to these principles, data should be assessed by its accessibility (as they must be open), by its format, and by its interoperability with other datasets. Tim Berners-Lee suggested a 5-star deployment scheme[20] for Linked Open Data: having the data on the web with open licensing; having structured data; use non-proprietary open formats; using URIs to point at the data; linking data with other data.

Hasnain & Rebholz-Schuhman (2018) compared both sets of principles and considered that the main objective of LOD principles is data interoperability, and FAIR principles aim at reusability. The scope of FAIR principles is broader insofar as they can be applied to non-data assets as well (e.g. codes, workflows, etc.). There are other significant differences: whereas LOD mandates open data, FAIR requires a stated license for access; a key element of LOD principles is URIs, when FAIR allows for a broader range of identifiers. Finally, neither LOD nor the FAIR principles suggest any specific standard, technology, or solution. Both constitute a high-level guide for data producers and publishers.

[18] https://www.coalition-s.org/

[19] https://www.w3.org/egov/wiki/Linked_Open_Data

[20] https://5stardata.info/en/

In conclusion, the understanding of both the broader and the technical background shows that the FAIRification of a publishing service takes place in a complex environment that opens to various possibilities to better define the objectives of the FAIRification, but also requires addressing some specific constraints.

## 3.2 FAIR-enabling components in scholarly publishing

Based on this landscape study, it seemed useful to delve deeper into some technical definitions that are crucial for the FAIRification of publishing systems. These technical aspects can be seen as FAIR-enabling components, although they raise specific challenges in the context of scholarly publishing. These technical definitions are closely related to OpenEdition's use cases aforementioned as they were part of the assessment phase. We present them separately beforehand, for they could easily apply to other publishing services.

### *Persistent Identifier (PID)*

The term identifier as used in the context of digital identification refers to a label, a sequence of characters, which gives a unique name to an entity. This entity can be of different types: a person (researchers, authors, contributors), a place (institution, organisation, laboratory, a set of geographical coordinates), or a thing (publication, dataset, software). Persistent means it is an ongoing, long-lasting reference to the digital resource.

Persistent identifier is, thus, a non-semantic string of characters identifying a single object. It must be globally unique, persistent, and resolvable. Uniqueness and persistence are also characteristics of other identifiers, but in the case of the PIDs, such characteristics should be understood in reference to the digital environment. The PID has indeed to be unique in the context of the World Wide Web, persistent even in the unstable digital context, and always resolvable for a human or automated agent. A PID is essentially the mechanism that allows separating the identifier from the resource's location, i.e., the URLs, thus ensuring persistence. Uniqueness and correct resolution of the PID are managed through a registry that is maintained by an authority.

There are different PID systems, they are usually managed by global agencies, often for a fee. There are technically no obstacles for a local organization to maintain its own PID system, but the organization's limited perimeter and/or sustainability would lower its authoritative quality. Some well-known PID systems for objects are Handle[21], ARK[22], and the

[21] http://www.handle.net/index.html

[22] https://n2t.net/e/ark_ids.html

DOIs[23] from distinct registration agencies[24]. The Handle system is robust and can be installed internally for a minimal cost. The ARK system comes with interesting features for the management of hierarchical relationships between identifiers, which could allow for an accurate handling of a documentary unit's different available formats. In the field of DOI registration agencies, Datacite[25] provides DOIs similar to the Crossref[26] ones but for a minor cost.

As the publishing sector is involved in wide dissemination activities for a long time, it is not new to global identification. Publishing services, indeed, already ensure the identification of its objects through the ISBNs and ISSNs. However, as we can see, and even considering the digital-specific identifiers like e-ISSNs, these do not correspond to the PID definition. Like any index number, such identifiers can only be part of a PID or its resolution link. Furthermore, the PIDs' management relies on various agencies, which offer a variety of services according to different terms and conditions. The accurate evaluation of each distinct PID system, of the specific cost/benefit balance, represents a challenge for which little guidance can be found[27]. It was, therefore, one of the main objectives of this detailed assessment to review and compare the main existing PID systems. Finally, as already mentioned, the choice of a PID system is partially a forced choice in the publishing context. The current PID systems offer limited options in a sector that transformed some of these options as practical standards for high quality publishing services. For open access public organizations, this aspect requires particular attention - and imagination.

*Licenses*

According to the Open Science Training Handbook[28], "license is a legal document that grants specific rights to the user to reuse and redistribute a material under some conditions. Any right that is not granted by default by the licensor through the license can be asked".

In the scientific context, to apply a license on a work (a paper, a dataset, or any other type of research output) permits the copyright holder to express the conditions under which the work can be accessed, cited, reused, modified, etc. In the open access environment licensing mainly refers to open licenses[29], such as the Creative Commons (CC) ones.

[23] https://www.doi.org/factsheets/DOIKeyFacts.html

[24] https://www.doi.org/registration_agencies.html

[25] https://datacite.org/

[26] https://www.crossref.org/

[27] One example is the deliverable "Persistent and Unique Identifiers" by CLARIN (Wittemburg, 2009).

[28] https://open-science-training-handbook.github.io/Open-Science-Training-Handbook_EN//02OpenScienceBasics/06OpenLicensingAndFileFormats.html

[29] Open Knowledge Foundation's definition is available at: https://opendefinition.org/.

For reusability purposes, the FAIR principles recommend providing, both for humans and machines, clear information about licensing. In the open access context, however, licensing mainly refers to open licenses[30], such as the Creative Commons (CC) licenses. Providing clear licensing information depends on the type of objects to which the license is applied and also on the existing regulations at a national level.

Under French law, and generally at a European level, there is no distinction between publications and data, but between intellectual work and information. Stérin (2018) explains that "data" does not exist as a legal object. It means that data, in itself, does not fall under a specific legal regime. The law only knows about personal data (whose use is strictly regulated) and public sector information, most of which is *a priori* freely accessible and reusable.

In the French context, we can resort to the Act for a Digital Republic (*Loi pour une république numérique*[31]) of 2016 to understand the status of research data. The Act determines an open status by default of information produced by administration units with more than 2500 agents. It also determines the free reuse (including commercial), with few exceptions (protection of rights belonging to third parties: intellectual property, privacy, confidentiality, and secrets). Therefore, research data are well subject to the principle of opening by default. France has defined by decree two possible licenses for such data. They are  the open license for the reuse of public information[32] and the Open Database Licence[33]. CC licenses, on the contrary, are not yet validated for these objects.

The by-default opening principle, however, does not apply to scholarly publications. Maurel (2018) explains a significant difference in the legal regime applicable to scholarly work and information. Scholarly publication falls under the category of intellectual works ("*oeuvres de l'esprit*", in French law): they are characterized by an original quality giving birth to authorship rights. Still conserving such rights, authors can agree to extend the possibility of reuse of their creations through the use of open licenses. CC licenses, for instance, are a well-spread standard for publications open licensing that offers various options to modulate the possibilities of reuse.

In the case of a publishing system, it appears that the first work to conduct is an accurate inventory of both intellectual works and information. Although the identification of intellectual work is easy for the textual contents, the publishing system handles and generates a wider range of content that has a less obvious status. It requires taking specific actions for

[30] Open Knowledge Foundation's definition is available at: https://opendefinition.org/.

[31] Act number 2016-1321, from Octobre 7th 2016, for a Digital Republic (Loi pour uneRépublique umérique): https://www.legifrance.gouv.fr/eli/loi/2016/10/7/ECFI1524250L/jo/texte.

[32] https://www.etalab.gouv.fr/wp-content/uploads/2017/04/ETALAB-Licence-Ouverte-v2.0.pdf.

[33] https://spdx.org/licenses/ODbL-1.0.html#licenseText.

any additional materials of third-party authors contained in the publications (images, drawings, etc.). On the contrary, the metadata mechanically generated, generally cannot be proved to be intellectual work. An exception might be the summary, which can be considered an intellectual work and requires, therefore, to establish agreements for the attribution of a liberal license to all the metadata (CC0, for instance). Like the descriptive metadata, the TEI digital mark-ups of published contents are not an intellectual work themselves; it is however possible to specify distinct licenses for different formats of the same work.

A second conclusion can be made from this legal context: the clarification of the licensing also implies interacting directly with the publications' right owners, the authors, and the publishers representing them. Further discussions are necessary, as well as specific legal expertise, to come to agreements about the licensing policies and options to adopt at the level of the organization[34].

*Authors' information management*

Handling bibliographical data implies being able to disambiguate legal or physical persons. Identification by name is often insufficient and it can become hard to distinguish homonyms. In addition, name changes may occur, which produces many ways of referring to an author, sometimes by initials or inverted forms. To address these challenges, it is possible to use authoritative registries, which, in the digital context, can correspond to a specific type of PIDs. The OpenEdition team collected information on three authoritative registries for persons' unambiguous identification: ORCID, Idref, and VIAF.

ORCID initiative represents a specific case insofar as it provides persistent identification for authors. However, the authors themselves provide the information, which is not curated. Each author can create his/her ORCID Id, a persistent identifier, and then link his/her publications to the ORCID id.

IdRef is a platform of the French Bibliographical Agency for Higher Education, the ABES. It aggregates different authority registries and provides a web interface, a triple-store as well as web services. IdRef is designed for collaboration: users, according to their rights, can modify records or report errors.

VIAF is a website that pools the resources of different libraries to provide a common and shared authority file. The VIAF data contains general information (nationality, working language, alternative spellings), the author's publications, co-contributors and publishers, links to the record in other repositories, and a history of the record. The data is available under the open license Open Data Commons Attribution license 1.0 (ODC-By).

[34] It is probably worthwhile noticing that the work conducted on licensing did not only rely on documentation, but also on direct consultation with one of the authors, namely L. Maurel.

## 4. OpenEdition's internal FAIR audit

The OpenEdition team produced an extensive internal report on its FAIR review. The objective of this paper is not to provide a complete summary of this report, but rather select the more relevant aspects of the methodology employed. For this reason, this paper may give more details about specific platforms: OpenEdition Books and OpenEdition Journals. Nevertheless, the OpenEdition FAIR review considered all the infrastructure's datasets, which have all undergone the FAIRification process.

The following sections describe this general process, referring to OpenEdition's services as an illustration.

### 4.1 FAIR assessment

*FAIR principles' contextualization*

The FAIR principles aim at increasing, both for humans and machines, the Findability, Accessibility, Interoperability, and Reusability of digital scholarly resources. It is necessary to transpose these general objectives to the specific context in which one performs the FAIR review. The 15 FAIR definitions and commentaries are therefore analyzed in the light of the publishing service practices, aims, and features.

At this first level of analysis, we can make two main observations. On the one hand, only a few FAIR principles seem difficult to apply in the publishing context. Such difficulty is mainly the case for the principle R1.2, which states that "(Meta)data are associated with detailed provenance". It is possible to interpret the provenance as the roles held by the publishers and the authors, but the process of creation of the published digital object is rarely described as the process of creating research data. On the other hand, for an open access publishing service that is natively digital and essentially focused on dissemination, many FAIR principles are already addressed, even if not extensively.

Findability and Accessibility are under the responsibility of the infrastructures rather than of the data producers. It is also the case for publishing services, especially open access ones. Each data must have a unique and persistent identifier (PID), this is a prerequisite for all the other principles. While for datasets a fully functional PID such as Handle can meet the expectations, the high-quality referencing expected by the publishing service's customers implies to use *de facto* standards like DOIs, which come at a financial, technical and human resources costs (the detailed assessments section will give more information on identifiers). For this reason, DOIs may not be used for all the data generated, thus limiting its extensive findability. Accessibility is one primary goal of open access publishing, with restricted access being the exception. The use of an open protocol such as HTTP(S) facilitates the access to the contents, but it also requires further developments to manage authentication and authorization in a more automated way.

For traditional editorial forms like books and journals, Interoperability can be reached through the use of interoperable standards both for the data (e.g., TEI, JATS) and the metadata (e.g., DublinCore, METS). However, for less traditional forms, like blogs and scientific events—as is the case of OpenEdition Hypothèses and Calenda—, interoperability is hindered by the lack of similar standards. It is noteworthy that interoperability should also consider community standards, which in our case could be either the publishing community or the SSH community (for example, disciplinary controlled vocabularies). In both cases, the recommendation to have these controlled vocabularies FAIR-compliant themselves require specific attention.

We can ensure Reusability when we do not presume which metadata is useful to whom and provide all the information available. It seems, however, difficult to identify in the publishing service, especially when it provides the tools for the datafication, what constitutes the raw data, and, as a consequence, the precise provenance trail. The question of a clear licensing is also challenging, given the variety of digital objects managed by the service and the distinct legal provisions applying to them. Furthermore, the information system has to make the licensing information available for an automated agent.

The FAIR principles contextualization, as we summarized, gives us an overview of the principles' specific expression in a publishing service and already gives indications on which areas will have to be surveyed more intensely.

### *Data definition*

In this specific context, the FAIR principles implementation seems highly dependent on the type of data considered. Therefore, the second step of the FAIR assessment consisted of the definition of the datasets to analyze. In the case of OpenEdition, the first series of datasets naturally relates with the four publishing platforms: OpenEdition Journals, OpenEdition books, Hypotheses, and Calenda. However, a publishing system generates and processes other datasets, which stem from added-value services or the information system monitoring. In the prospect of the full FAIRification of OpenEdition's data, it was decided to not exclude any dataset. First, because the FAIR assessment process could be used as a global assessment of the data and service provision of the organization. Second, because it conforms with the Open science goals of making any data FAIR, anticipating any potential use of any digital data. The datasets listing therefore included not only traditional publishing forms like journals and books, but also blogs and scientific announcements, and other potentially reusable datasets.

The simple listing of all these datasets with their main characteristics alone provides us some information regarding the current or the potential level of FAIRness of each dataset (Table 2).

Table 2. OpenEdition's main datasets selected for the FAIR review

| **Dataset** | **Type** | **Software** | **Schema*** | **Access**** | **Creator** | **Finality** |
|---|---|---|---|---|---|---|
| Journals | Journals<br>Articles<br>Others | Lodel | *Data:*<br>TEI<br>*Metadata:*<br>DC<br>METS<br>MARC<br>ONIX | *Data:*<br>HTML<br>PDF<br>ePub<br>*Metadata:*<br>OAI-PMH | Author<br>Publisher | Dissemination |
| Books | Monographs<br>Chapters<br>Others | Lodel | *Data:*<br>TEI<br>*Metadata:*<br>DC<br>METS<br>MARC | *Data:*<br>HTML<br>PDF<br>ePub<br>*Metadata:*<br>OAI-PMH | Author<br>Publisher | Dissemination |
| Hypotheses | Blogs/posts | Wordpress | *Metadata:*<br>DC | *Data:*<br>HTML<br>*Metadata:*<br>OAI-PMH | Author | Dissemination |
| Calenda | Announcements | Lodel | *Metadata:*<br>DC | *Data:*<br>HTML<br>*Metadata:*<br>OAI-PMH | Author<br>OpenEdition | Dissemination |
| Vocabulary | Terms | (Open Theso) | (Internal) | *Data:*<br>HTML | OpenEdition | Enrichment |
| Training corpus | Enriched TEI | | *Data:*<br>TEI | *Data:*<br>Github | OpenEdition | Enrichment |
| Metrics | Metrics | Matomo | | *Data:*<br>HTML | Matomo<br>OpenEdition | Monitoring |
| Catalogs | Detailed listings (books, journals, blogs) | | *Metadata:*<br>Kbart | *Data:*<br>HTML<br>CSV<br>TXT<br>XLS | Publisher<br>OpenEdition | Discovery |

* The "Schema" column lists schemas used both for data and metadata.

** The "Access" column lists access pathways used both for data and metadata.

Table 2 shows that the datasets are firstly defined by their access points (e.g., public platforms, internal interface) and by their object types. They are also, more precisely, defined by: the software used to manage the data, the schemas applied to the data and the metadata, and the format in which the data is available (see Annex 1 for details). For journals, books, and events, OpenEdition uses the home-built CMS Lodel already mentioned. The blogs are created with the CMS Wordpress.  A full-text TEI version is available for journals and books only. Metadata is available in the DublinCore and METS formats in different sets of the OAI-PMH repository. Additionally, MARC records are created for the libraries and ONIX books' metadata records for the bookshops. Finally, the role of OpenEdition in the production of such datasets also defines them.

While the models and the software solutions used for the data and metadata generation impact the findability and interoperability, the type and the creator can affect reusability because of the specific applicable open licenses.

*FAIR analytical review*

At the core of the FAIR assessment process lies the full FAIR analytical review of each dataset. Such analytical work is necessary to avoid a generic application of the FAIR principles. In fact, it helps identify the actions required towards FAIR.

The analytical review used a table comparable to the FAIR data maturity model developed within RDA (RDA, 2020), with a lesser level of detail and without specific indicators, but with more space for comments and appreciation. It seemed more appropriate to assess the FAIRness of the publishing system in a more comprehensive and graduated way. The level of analysis is the global dataset generated by the publishing system, comprising the data created (e.g., PDF or TEI files) and their related metadata (e.g., standard DublinCore or Wordpress metadata). It didn't consider the ingested data (e.g., .docx or .odt files), as these are not made findable or accessible, and therefore do not enter into the FAIR scope.

The analysis evaluates the global level of FAIRness of the documentary units within a dataset. For each dataset, the analytical table contains a short description: creator of the data, expressions of the data and of the metadata. The creator of the data can be the author, the publisher, or the OpenEdition's team. Expressions of the data, with reference to the FRBR model[35], cover the dataset's various formats and uses within the information system. For the metadata, the table specifies if it conforms to a standard or not and which one. The table displays, then, for each FAIR principle, the current FAIRness status of the dataset. It also shows the existent elements that allow the FAIRification (FAIR-enabling elements) and the ones that hinder FAIRification.

[35] https://en.wikipedia.org/wiki/Functional_Requirements_for_Bibliographic_Records

A selection of these tables is reported in Annexes 1 and 2. For instance, in the case of the journals' platform in Annex 1, in the line for principle F1 "(Meta)data are assigned a globally unique and persistent identifier", all the existing PIDs for the platforms' objects, even PIDs created by other organizations, are listed as FAIR implementations. The column "FAIR-enabling information" lists the information existing in the system which could be used to achieve F1, in this case the identifiers used in the OAI-PMH repository. The last column lists instead aspects that represent either a limitation of F1 (amount of objects actually having a PID) or a challenge for F1 (retrieval and integration of external PIDs).

In Annex 2, the analytical table allowed a clear FAIR assessment of OpenEdition's controlled vocabulary. The dataset description specifies the creation process of this vocabulary within the organization and its planned integration in a thesauri management tool. The three columns table gives a detailed evaluation of the current situation, the planned improvements, and the potential evolutions. The FAIR principles indeed gave a consistent analytical grid to assess the quality of the vocabulary in the prospect of its use, reuse and integration in the broader digital landscape. Whereas the vocabulary in its previous stage is used only internally, does not provide PIDs for the concepts, is searchable only through the platform's filters, and does not contain semantical structuration, the integration into the thesauri management tool will instead address all these challenges. In this case the FAIR assessment helped to validate a planned action with sound and coherent arguments.

It is worth noting that the tables represent an effort of documentation, which is in itself a FAIRification achievement. We report below the key challenges for FAIRification identified for the more relevant datasets in the OpenEdition's case[36].

- OpenEdition Journals:

The data types include articles, issues, and collections, among others such as reviews. All these types are considered as primary data, as OpenEdition's service does not comprise the editorial work. Not all the types receive a DOI, both for financial and technical reasons. The documentary units without DOI are only identified through the identifier of the OAI-PMH repository, which does not have all the functionalities of a PID (Wittenburg, 2009). Due to the absence of a dedicated registry for authors or the connection with an external database, most authors are not identified through a persistent identifier, except for a minority who are identified through ORCID. The core issue regarding accessibility comes from deleted records, which remain available for the harvesters in the OAI repository. The open licensing issue requires clarification due to the coexistence of some elements: external requirements, competing legal provisions, and distinct dissemination policies for the different formats.

[36] The challenges reported concern the status at the time of the review, a certain amount of them have been addressed since.

- OpenEdition Books

The data types include books, chapters, and collections, but also other types as bibliographies. Like for journals, all these types are considered as primary data, as OpenEdition's service does not comprise the editorial work. Regarding the persistent identification of digital objects and authors, the same observations made for journals are applicable. Furthermore, the books are enriched with controlled vocabularies that could be FAIRified (see below). Regarding reusability, the organization created a specific open license. The organization should still assess the validity of this license regarding the FAIR principles requirements.

- Hypotheses

The level of analysis is the post, consisting of content and related metadata created and managed through the Wordpress software. The blogs and posts of the platform respect only minimal FAIR requirements. In other words, the documentary units do not receive DOIs, the metadata is dependent on the capacity of the software used (WordPress), and the keywords added are available only in the software databases. However, part of the metadata generated is made available in the OAI-PMH repository. Open licensing is not mandatory; it is only recommended and left to the appreciation of the authors. The licensing information is, however, not integrated with the global information system.

- Calenda

This platform contains scientific events co-authored by the announcer and the OpenEdition team. Like in the case of Hypotheses.org, the platform's content only respects minimal FAIR requirements. The two main differences concern interoperability and reusability: Calenda platform uses a controlled vocabulary that can be connected to community vocabularies; the legal status of the contents is uncertain due to the co-authoring.

- Vocabularies

The shared OpenEdition Index is an internal controlled vocabulary of 188 terms with the translation available in various languages (DEU, POR, ENG, SPA, ITA, FRA) used to describe documentary units managed through the Lodel software. It lacks at the moment the qualities to be considered a FAIR vocabulary. Nevertheless, its integration into a thesauri management tool (OpenTheso[37]) will allow to: add PIDs (Handle or ARK) to the terms, manage the deleted records, add a semantic layer for hierarchical links (SKOS-RDF), and to enrich the vocabulary documentation.

- Training corpus

[37] Opentheso is a multilingual thesaurus manager developed by a CNRS research team, and supported and hosted by Huma-Num. More details at: https://opentheso.huma-num.fr/opentheso/.

The OpenEdition Lab produced tools to add new services to the various platforms (for example, Bilbo, a tool for the automated annotation of bibliographical references[38]). Some of these tools required the creation of annotated corpora for machine learning. The corpora are available on Github, most of them in TEI format, as they were created from OpenEdition's contents. The main challenge concerning FAIR is the possibility of reuse, which is for now limited to the Text and Data Mining exception granted by the French law.

Regarding the metrics datasets, although its FAIRification may also be important in the prospect of usage analysis, it is dependent on the software generating them, which limits both the assessment and the implementation of the FAIR principles. In the case of the catalog dataset, although it represents an important service for the users, it consists only of metadata which is fully standardized and does not pose major FAIR issues.

The full FAIR analytical review allowed therefore to list with accuracy the FAIR existing or potential components and the main challenges faced for each dataset. We can conclude that the overall FAIR maturity level of the OpenEdition publishing system is very uneven and hindered by contextual aspects and by the non-traditional publishing typologies. Furthermore, the analytical full review unveiled the main areas where we can improve the level of FAIRness and those for which we needed more detailed information to formulate more accurate recommendations.

### *FAIR synthetical review*

The full analytical review joined with the previous use case analysis led to specific synthetical assessments, which already prepared the way for the phase of recommendations. These specific assessments were the following: persistent identifiers, licensing, and author's information management.

#### *Persistent Identifiers*

OpenEdition has implemented the Crossref DOIs[39] for some of its contents. These PIDs imply, however, some limitations. They are not applied to all the object types of OpenEdition's platforms, for example the Calenda's scientific announcements. Therefore, only books and journals documentary units receive a PID: 90% and 45% of the documentary units for books and journals, respectively. Documentary units without DOIs can be reports, editorials, chronicles, or archaeological notices. This limited implementation of Crossref DOIs is partially due to financial aspects (the estimated cost of DOIs for all documentary units amounts to 27,000 USD). However, Crossref DOIs implementation also implies

[38] https://www.openedition.org/9202?lang=en

[39] https://www.crossref.org/

technical challenges. In the open access context, publications can be accessible via many platforms, which implies managing the multiple resolution links accordingly. The existing solution for such management is highly dependent on the coordination with the primary DOI creator and uneasy to implement in a straightforward way.

In all the other cases, as mentioned before, the documentary units are identified internally according to this syntax: Platform*Sitename*Lodel_Id. A similar syntax is used in the OAI-PMH repository. The syntax proved rather efficient to manage URLs changes (e.g., https://remi.revues.org/7777 and http://journals.openedition.org/remi/7777 both redirect correctly after the platform's name changed). Nevertheless, contrary to the PID definition, this syntax does not separate the identification from the location and does not fully ensure the persistence. Furthermore, the information system does not correctly manage the deleted records: the identifiers (DOIs or internal) remain available in the OAI-PMH repository with no information about the deletion for the harvesters.

To increase the coverage in PIDs and improve the information system, the OpenEdition team thus reviewed the specifications, features, and cost of various PID systems: Handle[40], ARK[41], PURL[42], and the DOIs[43] of distinct registration agencies[44]. The Handle system is robust and can be installed internally for a minimal cost; it is the system underlying the DOIs' systems, even if with less features, and it is already used in OpenEdition's environment (Isidore platform, OpenTheso). The ARK system comes with interesting features for the management of hierarchical relationships between identifiers, which could allow for an accurate handling of a documentary unit's different available formats. In the field of DOI registration agencies, although often used for datasets, Datacite[45] provides DOIs similar to Crossref DOIs for a minor cost and with a metadata schema that fits OpenEdition's needs.

### *Licensing*

Currently, at OpenEdition, the modalities of reuse are defined in different and not always consistent ways. They are mainly defined by contractual documents: the Terms and Conditions of Use and the General Conditions for Commercial dissemination. Modalities of reuse can differ depending on the access mode (full open access or open access limited to HTML version). In some cases, the modalities of reuse are also defined by a specific original license (the OpenEdition license), or by the declaration of a CC[46] license. However, there is no general policy for CC licensing, which can differ within a platform or from one platform

[40] http://www.handle.net/index.html
[41] https://n2t.net/e/ark_ids.html
[42] https://sites.google.com/site/persistenturls/
[43] https://www.doi.org/factsheets/DOIKeyFacts.html
[44] https://www.doi.org/registration_agencies.html
[45] https://datacite.org/
[46] https://creativecommons.org/

to another. CC licenses generally appear on the published contents or web pages instead of being integrated into the information system.

For journals, the default license is defined for all publishers with a few exceptions. In fact, in 2016, the new requirements by DOAJ[47] resulted in several journals changing their default license to a CC license. This change was applied retroactively to all the journals and the validity of these licenses might be therefore questionable. For books, a license (CC or OpenEdition for Books) can be defined at the book level or the publisher level. Approximately 1300 over 10000 books indicate a license, but the management of that information in the system is uneven. There are no license specifications for publications on Calenda. The authors of the announcements are not clearly defined, as the Calenda team reworks the ad (rewording, layout, addition of keywords), they are also the author. However, for the same reason, setting up a general licensing for this platform does not imply greater risks. Hypotheses team recommends the use of Creative Commons licenses. This information is visible on the website of the blog but not retrieved in the OpenEdition system.

Finally, besides the publications licensing, the OpenEdition's 2020 Terms and Conditions of Use[48] specify that the organization may carry out text mining and data processing on publications and that a researcher may request access to OpenEdition's data. Such TDM usage is indeed already in place within OpenEdition's laboratory for the creation of annotated corpora. Like in the case of the ROBOH corpus[49]. However, even if the corpus is freely accessible under an open license, the possibilities of reuse and/or republication remain uncertain. It is a more general challenge for the TDM rights management: the law acknowledges the TDM exception for scientific purposes, but gives few provisions about the republication possibilities.

### *Authors' information management*

The author's information management in OpenEdition lacks consistency and of connections with external registries (see the section: *FAIR-enabling components*). No internal database aggregates all the authors nor serves as the basis for a general index of authors. As a result, the information on authors is scattered in the information system. The authors' information is indeed attached to the metadata of the documentary units. The information is therefore manageable to some extent in the system: it is available for the various objects' expressions (TEI, METS, DublinCore, MARC); it is searchable on the web interface. OpenEdition also implemented the possibility for the authors to connect directly to their ORCID account and link OpenEdition's publications that match their name. Although technically satisfying, this solution has some limitations due to the human errors it can imply.

[47] Directory of Open Access Journals (DOAJ): https://doaj.org/

[48] https://www.openedition.org/31127?file=1

[49] Review Of Books On Hypotheses (ROBOH): https://github.com/OpenEdition/roboh.

The synthetical assessments were the last step of this progressive assessment phase. They gave final details and leads to establish a list of recommendations, classified according to their priority, regarding both the FAIR principles and the service improvement.

## 4.2 Recommendations

The final phase of the full FAIR review consists in assigning "relative priorities to recommendations" and associating "actions to the top-priority recommendations" (Koers et al., 2020b). We present hereafter the recommendations that were validated within OpenEdition through the FAIR assessment process. For the top-priorities of the organization, the recommendations defined an action plan. The recommendations also listed a number of further actions which would improve the FAIRness of the publishing system. These recommendations are related to OpenEdition's specific case, and they should serve only as an illustration of the results that can emerge from the overall methodology.

*Action plan*

- Persistent Identifiers

The objective for OpenEdition is to attribute PIDs for all the published contents and more generally for all types of data, in particular by maintaining a database connecting PIDs and metadata, even after contents' records have been deleted.

The use of DOIs as the PID system for all the resources represent a technical challenge: there are at the moment no satisfying solutions to manage the additional DOIs of journals and books published on other platforms. Furthermore, the attribution of Crossref DOIs for the documents of all the platforms would represent a significant financial cost. Other registration agencies (such as Datacite) would however allow for a more economical solution.

Therefore, although keeping in use the Crossref DOIs and their reference linking services, a more flexible PID system, like handle.net or ARK, can be used for all the data generated by the publishing system. The final choice is to implement Handles as the by-default identifier: they are technically close to the DOIs and already in use in the OpenEdition's environment (Huma-Num). The implementation of Handles can be achieved internally or outsourced. Putting in relation in a database, the PID, the URL and the metadata would allow the provision of the metadata of a deleted record. With a minimal financial and technical cost, the Handles should therefore allow for a better management of the deleted record's information, both for the organization and for external services.

The recommendation is therefore altogether to: implement Handles for all the data of the system; keep the Crossref DOIs where they exist; expand the coverage of DOIs through Datacite DOIs.

- Licensing

In OpenEdition's context, the objective is to attribute to all the content licenses stating clearly the possibilities of reuse. A distinction has to be made between the contents considered as information, and those considered as intellectual works.

Information includes any data produced by the public sector. In the case of OpenEdition, this type of objects refers to: metadata of the publications, the metrics, the data of the OpenEdition laboratory. This has particular importance for the open data project of OpenEdition. Provided that an exhaustive list of this public information is established, and GDPR[50] requirements for personal data are respected, they will be open by default and will have to select the two open licenses accepted under the French law. In the metadata, as the summary can be considered as an intellectual work, specific agreements with the publishers should be established in order to apply the most liberal licensing to the metadata. The application of CC-0 license on metadata, whenever possible, conforms indeed with recommendations and practices at the European level, and corresponds to OpenEdition's objectives of broad dissemination.

The published contents of the four platforms all fall under the category of intellectual works. The recommendation here is two-fold: establish a policy at the level of the organization; accompany the publishers and authors in the adoption of open licenses. The recommended policy is to adopt Creative Commons licenses, for they are well-spread and allow for persistent expression in the metadata. A CC license by default should be defined in the contracts with the publishers, allowing well-defined opt-out possibilities. The general policy may vary from one platform to another in terms of type (e.g. CC BY or CC NC) and granularity (e.g. blog and/or post). In the case of books and journals, it may also vary from one format to another, in order to conform with the contracts signed with the publishers (restricted access formats) and to define the use of specific formats (especially the TEI version). Specific training and support actions are planned to facilitate the publishers' and authors' engagement.

Additionally, the information about licensing should be better integrated with the information system. In the case of the TEI version, additional developments have to be planned in order to ensure automated authentication and authorization processes.

[50] General Data Protection Regulation of the European Union: https://gdpr-info.eu/.

*Further steps*

- Author’s information management

The information system should be updated in order to have the capacity to manage structured information about the authors. The authors’ database could then be linked with external authoritative registries (e.g. Idref). This would make it also possible to better specify the distinct roles of the authors of the contents.

- Controlled vocabularies

The recommendation concerning the OpenEdition shared vocabulary is to accurately describe its provenance and document its content. The use of Opentheso will notably increase the FAIRness of the vocabulary: PIDs for the terms and semantic relationships (hierarchy), thanks to the expression in SKOS. It thus becomes possible to envision alignments with other widely used controlled vocabularies (LCSH, EUROVOC, RAMEAU, etc.).

- Machine-actionability

This recommendation in our case mainly relates with the accessibility to the contents by the machines. Although the system uses only standard and open protocols for access (TCP/IP, HTTP, and OAI-PMH), the authentication and authorization are not directly managed by the protocols. Various leads are being explored concerning the integration of an Authentication and Authorization Interface (AAI) and HTTP mechanisms of content negotiation to access specific contents or formats of these contents.

- Digital Management Plan

As a continuation and an improvement of this documentation effort, a Data Management Plan of the entire publishing system is also recommended. The FAIR analytical review gave indeed the main elements to start a full description of the general data ingestion, generation, and delivery.

# 5. Lessons learned and perspectives

## 5.1 Lessons learned

Although it first appeared as a research data management tool and a part of the broader open science environment, the FAIR principles offer a consistent set of criteria also for the assessment of a publishing service. The work conducted within OpenEdition showed indeed that it was possible to assess the general quality of the service in terms of data and metadata management and provision. The work required however various adjustments with respect to the FAIRification of research data. First of all, it needed a landscape study which helped to connect the FAIR principles with other notions either already connected with publishing practices or simply better known. The assessment itself had to take into account, at the same time, the mixed nature of publishing platforms, both a service and a data repository, and the complexity of their workflows. To do so, it created a tool that allowed for more flexibility in the analysis. The analytical table did not provide the accurate indicators of the FAIR maturity model and remained at a high-level analysis, but it proved efficient to have a global overview of the datasets FAIRness. In fact, the decision to not omit any dataset from the analysis allowed to identify the most crucial FAIR issues throughout the system that the synthetical reviews would clarify, and to have the general view that would help to clearly assess the priorities. As mentioned before, the FAIR principles and the electronic publishing goals converge in many aspects, and even more so, the FAIR assessment provides a useful assessment of the general information system. The FAIR principles, however, do not cover all the requirements for a publishing platform (e.g., long-term archiving, version management), and OpenEdition's recommendations are mainly valid for this specific organization, but the overall FAIRification methodology, we believe, can be used by other publishing services as well.

Another important component of the FAIRification process should nevertheless complete this feedback. It concerns the organizational aspects, including both management and funding. In the case of OpenEdition, the first step was the set-up of a dedicated task force consisting of five people: three from the Data management department, one from the R&D department, and one from the International department. The composition of the task force obviously illustrates the centrality of data in a FAIRfication process, but also its complexity, with the additional perspectives of innovation and internationalization. However, the complexity goes even further. Internally, the assessment phase, especially the synthetical assessment, also collected inputs from the departments dedicated to the service delivery, i.e. the teams managing the platforms. Externally, the collection of information was not only based on documentation, but also on direct consultation with a legal expert. Even more so, the implementation of FAIR licensing plans to incorporate legal expertise into its process. Furthermore, the adoption of FAIR licensing practices will have to include the customers of the service, in our case the publishers and authors, through supportive actions consisting of a dedicated engagement program and a dedicated hiring.

All the above shows that the FAIRfication of a service implies more than local technical improvements or FAIR-scoring evaluation. In fact, the FAIRification process may have for the organization an additional cost in terms of financial and human resources. In the case of OpenEdition, the recommendations of the task force led therefore to the preparation of a project dedicated to the implementation of FAIR identifiers, licensing, and publishing standard formats, which obtained funding in 2020 through the French national call launched by the Fond National pour la Science ouverte (FNSO)[51].

## 5.2 A toolkit for FAIR publishing services

The FAIR review conducted by OpenEdition allows to gather the main elements of a toolkit for the FAIRification of publishing systems. Firstly, it provides a general framework, distinguishing the phases of the review and their specific steps. The toolkit should, in the same way, contain the general information and documentation necessary for the preparatory work (preparation phase), offer FAIR-assessment tools adapted to publishing systems (assessment phase), and provide guidelines about implementation strategies proper to academic publishing services (recommendation phase). Secondly, the FAIR review of OpenEdition can enrich the toolkit with detailed examples about a variety of challenges and use cases typical of publishing systems. Thirdly, the toolkit can reproduce the process of OpenEdition's FAIRification, which moved progressively from the more general to the more specific aspects, still taking into account the priorities of a publishing service.

The final toolkit should also, however, improve or further the work done at OpenEdition, either on specific or general aspects. Additional information should be given regarding metadata and publishing standards (e.g., JATS). The section about the "use cases" should be reshaped in order to give more accurate guidance for a thorough risks/benefits analysis prior to the FAIRification. The recommendation to establish a Data Management Plan should be mentioned as one of the first steps for achieving a FAIR-by-design data creation process. Generally, the toolkit should also support the process towards an increased machine-readability of the metadata and the data, such as the FAIRification of concepts within the content (Velterop, 2020). The technical readiness and capacity of publishers, especially in the open access context, can highly vary, and the toolkit allows for a modular and progressive approach of the FAIRification for these different situations. However, the final toolkit should address more specifically aspects related to a better connection between publications and data, and those related to the FAIR metrics implementation.

## Conclusion

FAIR principles are generic, but their implementation is contextual. It is particularly true in the case of a service that deals with a variety of objects and takes place in a complex environment. As we can see from the above, even for an open access publishing service focused on broad dissemination and reuse, the actual level of FAIRness, when considered

[51] To date (2021), the funding has already allowed OpenEdition to address the recommendations concerning licensing and to start the implementation of the new PIDs' policy.

thoroughly, still remains uneven. The FAIRification of a publishing service requires taking actions related both to the sustainability of the information system and to the quality of the service for the users. The specific mix of intellectual works and information, scientific and industrial standards, traditional and non-traditional editorial forms, describes a complexity that the FAIRification has to address. Such complexity determines a process where specific steps and priorities are identified.More generally, as a process, the FAIRification is not a one-stand action, and the implementation of FAIR principles has also to consider a long-term perspective by fully integrating the principles into the service's general management.

## References

Hasnain, Ali, and Dietrich Rebholz-Schuhmann. 2018. “Assessing FAIR Data Principles Against the 5-Star Open Data Principles.” In *The Semantic Web: ESWC 2018 Satellite Events*, edited by Aldo Gangemi, Anna Lisa Gentile, Andrea Giovanni Nuzzolese, Sebastian Rudolph, Maria Maleshkova, Heiko Paulheim, Jeff Z Pan, and Mehwish Alam, 469–77. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-98192-5_60.

Koers, Hylke, Morane Gruenpeter, Patricia Herterich, Rob Hooft, Sarah Jones, Jessica Parland-von Essen, and Christine Staiger. 2020a. “Assessment Report on ‘FAIRness of Services,’” February. https://doi.org/10.5281/zenodo.3688762.

Koers, Hylke, Daniel Bangert, Emilie Hermans, René van Horik, Maaike de Jong, and Mustapha Mokrane. 2020b. “Recommendations for Services in a FAIR Data Ecosystem.” *Patterns* 1 (5). https://doi.org/10.1016/j.patter.2020.100058.

Maurel, Lionel. 2018. “La Réutilisation Des Données de La Recherche Après La Loi Pour Une République Numérique.” In *La Diffusion Numérique Des Données En SHS - Guide de Bonnes Pratiques Éthiques et Juridiques*. Presses Universitaires de Provence. https://hal.archives-ouvertes.fr/hal-01908766.

Research Data Alliance FAIR Data Maturity Model Working Group. 2020. *FAIR Data Maturity Model: Specification and Guidelines*, 2020 https://doi.org/10.15497/RDA00050

Stérin, Anne-Laure. 2018. *Diffuser des données de la recherche dans le respect du droit et de l’éthique*. Presses universitaires de Provence. https://doi.org/10/document.

Velterop, Jan, and Erik Schultes. 2020. “An Academic Publishers’ GO FAIR Implementation Network (APIN).” *Information Services & Use* 40 (4): 333–41. https://doi.org/10.3233/ISU-200102.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Wittenburg, Peter. 2009. “Persistent and Unique Identifiers”. *CLARIN*, edited by Daan Broeder, Malte Dreyer, Marc Kemps-Snijders, Andreas Witt, Marc Kupietz, Peter Wittenburg. https://office.clarin.eu/pp/D2R-2b.pdf.

## Annexes

Annex 1: FAIR analytical review example: OpenEdition Journals

| FAIR review of OpenEdition journal data | | | |
|---|---|---|---|
| **Data summary** | | | |
| **Data sources** | Data produced through Lodel by publishers and users<br>Can be updated (not fully controlled by the organization)<br>Documentary units' distinct levels: text, issue and collection levels | | |
| **Data expressions** | Raw data: Lodel database (as used for the HTML expression)<br>Other expressions: TEI OpenEdition, PDF, ePUB<br>Metadata | | |
| **Commentary** | Different properties depending on the type (proper to Lodel software):<br>- Volume contains: Publications (issues, columns, annual columns); Documentary unit contains texts,<br>- Different types of texts (article, column, editorial, review, ...),<br>- Annexed files types can contain data (xls, csv, sound, image, video files),<br>Not all the different types are available in all the different expressions (TEI, pdf, epub)<br>Question: Should the review consider the types that don't correspond to specific content (subpart, section, site, directory, etc.)? | | |
| | **FAIR implementations** | **FAIR**<br>**enabling information** | **FAIR limitations** |
| **Findable** | | | |

| | | | |
|---|---|---|---|
| F1. (Meta)data are assigned a globally unique and persistent identifier | Objects:<br>- DOI (prefix 10.4000): available only for some data (depending on the types and publishers' wishes)<br>- Handles generated by Isidore harvesting platform (not retrieved by OE)<br>Persons: a few Orcid<br>Organizations: a few IDs from Crossref Funding registry. | - OAI identifiers exist for all documentary units but are not PIDs<br>- All documentary units are identified in the information system though the concatenation: Platform+SiteName+ID | Objects:<br>- Some data without any PID<br>- DOIs may exist for data published on another platform that we do not retrieve.<br>- Handles assigned by Isidore are not retrieved.<br>Persons:<br>Contributors are not linked to registries. |
| F2. Data are described with rich metadata (defined by R1 below) | - Metadata available in the OAI-PMH repository (could be richer)<br>- Formats DublinCore, DublinCoreTerms, METS | Rich metadata is available; could be extensively integrated in the OAI repository.<br>In OAI, metadata available only for certain types (subpart, heading, and news are missing) | |
| F3. Metadata clearly and explicitly include the identifier of the data they describe | In the OAI repository:<br>- ID OAI<br>- DOI when available | | Some data without any PID (see F1) |

| | | | |
|---|---|---|---|
| F4. (Meta)data are registered or indexed in a searchable resource | OpenEdition Search interface (search.openedition.org):<br>- only a selection of data is available (some types are excluded),<br>(Meta)data is also searchable in other directories (e.g. Isidore harvests OE’s OAI repository) | No public API available yet, but all the information is available through the search software (SolR) | Metadata are not complete. |
| **Accessible** | | | |
| A1. (Meta)data are retrievable by their identifier using a standardised communications protocol | HTML: accessible via the DOI<br>Metadata: accessible via the OAI identifier | | Some data without any PID (see F1) |
| A1.1 The protocol is open, free, and universally implementable | HTTP for the data<br>OAI-PMH for the metadata | | |
| A1.2 The protocol allows for an authentication and authorisation procedure, where necessary | All protocols are open, but not all allow for authentication.<br>Protocol used for restricted access contents: | | Lack of a tool dedicated to the management of authentication and authorization processes. |

| | | | |
|---|---|---|---|
| | - TCP/IP for contents requiring authentication (TEI version’s case)<br>Other protocols used where authentication is not required:<br>- HTTP for open access contents<br>- OAI-PMH for the metadata | | |
| A2. Metadata are accessible, even when the data are no longer available | No | | No records for the deleted data. |
| **Interoperable** | | | |
| I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. | TEI, DC, METS | | No semantic layer is implemented. |
| I2. (Meta)data use vocabularies that follow FAIR principles | In some journals, use of disciplinary controlled vocabularies (e.g. French Pactols). | Some disciplinary controlled vocabularies (JEL, GeographieUN) could be integrated with thesaurus management tools | For most of the journals, no controlled vocabulary is used. |
| I3. (Meta)data include qualified | In OAI repository:<br>- is part of | - Citation and Cited-by available but not disseminated | Link with translations not recorded in the OAI repository |

| | | | |
|---|---|---|---|
| references to other (meta)data | - relation with OpenAIRE accessright field<br>Some links with translations | - on-going project: OE Review of Books | |
| **Reusable** | | | |
| R1.1. (Meta)data are released with a clear and accessible data usage license | Licenses are defined by journals and not by documentary units, except in a few cases.<br>The license is not defined according to the different expressions, the same license applies for all. | License should be distinct for each expressions and the information be added to the database and the TEI | No clear provision to allow for the text and data mining exception (acknowledged by French law "Loi pour une république numérique")<br>The license applied to the documents is not always clear.<br>The license has sometimes been declared by the journal retroactively, and has therefore uncertain value. |
| R1.2. (Meta)data are associated with detailed provenance | Internal creation process of the data is not described (can be created through Lodel, outsourced digitization, etc.) | | |

| | | | |
|---|---|---|---|
| R1.3. (Meta)data meet domain-relevant community standards | I1: (meta)data meet community standards for textual contents, including TEI.<br>I2: Fewer (meta)data meet disciplinary communities standards. | Semantic expression of the OpenEdition's controlled vocabulary (in SKOS) could help to connect with other SSH disciplinary vocabularies. | |

## Annex 2: FAIR analytical review example: OpenEdition controlled vocabulary

| FAIR review of OpenEdition/Calenda shared vocabulary | | | |
|---|---|---|---|
| **Data summary** | | | |
| **Data sources** | Controlled vocabulary developed internally: OE team and Scientific Board<br>SSH focused 188 entries covering topics, geographic areas, and periods of time.<br>Aligned with broad categories from: CAIRN, Érudit, HAL. | | |
| **Data expressions** | Used for all Calenda platform's contents.<br>Partially used by other platforms and services.<br>Facet of the search interface.<br>Terms available in: DEU, POR, ENG, ESP, ITA, FRA | | |
| **Commentary** | The vocabulary is currently being integrated with a thesaurus management tool: OpenTheso. This new implementation constitutes the FAIR enabling information described below. | | |
| | **FAIR implementations** | **FAIR enabling information** | **FAIR limitations** |
| **Findable** | | | |
| F1. (Meta)data are assigned a globally unique and persistent identifier | No | Assignment of PIDs to the terms (ARK or Handle via OpenTheso) | |
| F2. Data are described with rich metadata (defined by R1 below) | No, only a correspondence between an alphanumeric code | | Creation of descriptions for each entry, similar to Clarivate's "Scope Notes". |

| | | | |
|---|---|---|---|
| | and the terms in the various languages. | | |
| F3. Metadata clearly and explicitly include the identifier of the data they describe | N/A | OK | |
| F4. (Meta)data are registered or indexed in a searchable resource | Possibility on the interface to search by the "themes" corresponding to the vocabulary entries. | Terms will be searchable via OpenTheso. | |
| **Accessible** | | | |
| A1. (Meta)data are retrievable by their identifier using a standardised communications protocol | N/A (no PID) | On OpenTheso: access via the identifier through the web interface or the REST API. | |
| A1.1 The protocol is open, free, and universally implementable | N/A | OK (HTTP / REST) | |
| A1.2 The protocol allows for an authentication and authorisation procedure, where necessary | N/A | Authentication managed through the web interface not directly the protocol (RFC 2617) | Authentication by the protocol |
| A2. Metadata are accessible, even when the data are no longer available | No | Identifiers of a deleted resource are deprecated. | |

| | | | |
|---|---|---|---|
| **Interoperable** | | | |
| I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. | No | Structured representation with SKOS-RDF / JSON-LD / Turtle | |
| I2. (Meta)data use vocabularies that follow FAIR principles | No | N/A | N/A |
| I3. (Meta)data include qualified references to other (meta)data | No | Possible alignments between ontologies, semantic and hierarchical links within an ontology. | Alignments with external standard vocabularies |
| **Reusable** | | | |
| R1.1. (Meta)data are released with a clear and accessible data usage license | No | | License missing. |
| R1.2. (Meta)data are associated with detailed provenance | No | | Description of the vocabulary creation and update processes. |
| R1.3. (Meta)data meet domain-relevant community standards | Partially | | Alignments with external standard SSH vocabularies |