FAIRifying a scholarly publishing service: Elements for a toolkit based on the OpenEdition's FAIR review

Authors: Karla Avanço[1], Arnaud Gingold[2]

Abstract:
The FAIR principles (Findability, Accessibility, Interoperability, and Reusability) constitute a guide whose aim is to improve the management of digital scholarly resources. Nevertheless, the literature regarding data services other than data repositories is still scarce. OpenEdition is a digital infrastructure for open scholarly communication in the Social Sciences and Humanities (SSH) that carried out an internal full review to assess the degree of FAIRness of its activities. The objective of this paper is to present the methodology employed by OpenEdition's team and the recommendations for the FAIRification of a publishing system, and hence, the elements for the FAIR Publishing Toolkit. The FAIR review was conducted in three main phases: preparation, assessment, and result phase, which listed the recommendations for the FAIR principles implementation. The preparation phase gathered the available information to define the perimeter of the FAIR review. It comprised two steps: the landscape study and the exam of actual use cases. The assessment phase contextualized the FAIR principles according to the scholarly publishing context, defined the datasets to be analyzed, carried out a FAIR maturity review per dataset, and analyzed the state of the art of some important FAIR-related elements. The result phase produced the recommendations, organized as priorities and extended objectives. The priority recommendations regard persistent identifiers and licensing policies. The extended objectives focus on authors' information management, controlled vocabularies, machine-actionability, and Digital Management Plans.

[1] CNRS, OpenEdition, ORCID: https://orcid.org/0000-0001-8784-7754
[2] CNRS, OPERAS

Table of Contents

Introduction **3**

Methodology: Addressing the publishing specificities **4**

The service provision of a publishing system 4

FAIR review methodology: A three phases project 5

Preparation phase: Context and Challenges **7**

Landscape study **7**

Use cases **9**

Assessment phase: The FAIR principles as a grid **10**

Contextualized FAIR principles **10**

Data definition **11**

FAIR maturity review **12**

Detailed assessments **14**

Persistent Identifiers 14

Licensing 16

Authors' information management 18

Result phase: Recommendations **19**

Priorities **19**

Extended objectives **20**

Elements for a FAIR publishing toolkit **21**

Conclusion **21**

References **23**

Annexes **24**

Annexe 1: Detailed FAIR principles **24**

Annexe 2: FAIR analytical review example: OpenEdition Journals **25**

Annexe 3: FAIR analytical review example: OpenEdition controlled vocabulary **30**

Acronyms and Abbreviations **33**

Recommended Reading **35**

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.

The initiatives for the FAIR principles adoption 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 can not 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 system. Although publishing systems tend to converge more and more with research data repositories, they still have specificities concerning data types, technical standards, workflows, and legal statuses. The main challenge here is to achieve both the primary goal of publishing systems—the dissemination—and their additional mission regarding machine-readable data provision—the *datafication*. These two aspects are not contradictory, and the FAIR principles precisely facilitate their convergence. However, in specific contexts, convergence requires more effort and attention. In the context of small and medium publishers, often open access publishers, the focus is mainly on dissemination, with little awareness or means to fully address datafication challenges. In the context of Social Sciences and Humanities (SSH), the research output is often a publication, which can require additional work to make this output available also as research data, i.e., to make it fully findable, accessible, interoperable, and reusable. The FAIRification of publishing systems implies, therefore, some adjustments of the principles to these specific contexts.

OpenEdition[3] represents an interesting use case in this prospect. It is a French digital infrastructure for open scholarly communication in the SSH domain that brings together four complementary platforms focused on journals (OpenEdition Journals[4]), book series (OpenEdition Books[5]), research blogs (Hypotheses[6]), and academic events (Calenda[7]).

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

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

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

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

[7] https://calenda.org/

Offering publishing and standardization services to SSH researchers and open access publishers, OpenEdition operates both as a dissemination and a datafication unit for its community.

In 2019, OpenEdition carried out an internal full review to assess the degree of FAIRness of its data. To do so, the organization established a task force that, after such evaluation, recommended some specific actions for the application of the FAIR principles. It was the birth of an on-the-job yet reasoned and efficient methodology. This paper draws from the extensive work done by OpenEdition's team.

This assessment process represents the first step into developing a FAIR Publishing Toolkit that will offer guidance for the FAIRification of publishing services. As FAIR principles are applied on a specific set of objects (publications, corpora, metrics) and by a distinctive actor (publication services provider), this toolkit targets structures similar to OpenEdition, that is, other publishing platforms. Presenting the methodology used at OpenEdition may allow other publication systems to ask the same questions and, hopefully, apply similar solutions.

This paper describes the main components of the methodology, reporting on OpenEdition's specific FAIR review as an illustration. Such components will provide the structure and content of the envisioned toolkit for the FAIRification of publishing systems: the FAIR Publishing Toolkit.

Methodology: Addressing the publishing specificities

The service provision of a publishing system

The first step to assess the degree of FAIRness of a publication service is to identify the particularities of this type of service because the FAIR principles need to be adapted to the context where they are going to be implemented.

According to the FitSM standard[8], a service is a "way to provide value to customers through bringing about results that they want to achieve" (FitSM, 2016), with the value being essentially the combination of utility and warranty. This simple definition allows to characterize the specificities of a publishing system in terms of value, customers, and expected results. As an information technology service, a publishing system mainly consists of storing digital data and delivering it with added value to its customers.

Customers are an "organisation or part of an organisation that commissions a service provider in order to receive one or more services" (FitSM, 2016). In fact, a service is not only a service for the end-user. The delivery of high-quality data to the platform's end-users, here the readers, is only the final stage of a process that binds the service provider to its primary customers. On the one hand, a publishing service firstly addresses the publishers' needs and, through the publishers, those of the authors. On the other hand, when operating in the academic context, a publishing service is also connected to the research community, through university libraries or important online catalogs. Such an environment of customers and users defines the added value expected from a publishing service. Dissemination is the first goal,

[8] FitSM is a light-weight free standard for information technology services management: https://www.fitsm.eu/.

with its proper, and often not aligned, standards, and with either poor relationship with research data standards (such as domain controlled vocabularies) or limitations to machine-actionability (as is the case of PIDs resolving on a landing page containing a non-typed link to a PDF document). Moreover, the publication of intellectual works comes with specific legal constraints that the FAIRification has to consider when addressing Accessibility and Reusability.

In many cases, however, the publishing service does not provide only wide dissemination (and therefore wide Findability and Accessibility) to its customers; it also operates as a datafication unit. In these cases, the data and its metadata are automatically accessible through open protocols, and the metadata is structured according to widely used standards. In fact, beyond the simple storage of the data and normalization of the metadata, a publishing system can also directly contribute to the datafication and FAIRification of the data by making these contents appropriately available for the digital environment. In the case of OpenEdition, for instance, part of the service consists of expressing the textual content in the interoperable XML format of the Text Encoding Initiative (TEI) for books and journals. In the prospect of FAIR, however, the Reusability of this expression also has to consider the specific legal status of intellectual works.

These few considerations show that the publishing services have a major role to play in the prospect of FAIRification, provided that their specificities as a service are taken into consideration.

FAIR review methodology: A three phases project

Given the above, the FAIRification process of a publishing service cannot be referred only to certification frameworks for data repositories, such as CoreTrustSeal[9], even if those would prove efficient FAIRification tools in a further stage. Although the FAIR principles were firstly designed for data, and the core activities of a publishing system consist in managing data, a FAIR review of a publishing system strongly connects the data management and the service provision aspects. This two-level analysis appeared in the various stages of OpenEdition's FAIR review. In lack of really appropriate frameworks, OpenEdition did not set up an extensive methodology for its FAIR review from the start but built it along the way instead. Nevertheless, considering the results, we believe that the methodology was efficient and can be better formalized for further reuse.

Some components of this methodology are common to many other assessment projects, for instance, the landscape study, the use case analysis, or the tasks' prioritization. Two aspects, however, characterize this methodology: the specific combination and articulation of the various activities and the use of the FAIR principles as an analytical grid. In fact, like 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).

At a general level, and in a rather traditional way, the FAIR review was performed in three main phases: preparation, assessment, and result. The first two phases contained

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

specific sections and resulted in a set of recommendations. Table 1 synthesizes the phases employed in the methodology.

Table 1. The three main phases and the distinct steps of the FAIR review

PHASES	**STEPS**			
Preparation	Landscape study		Use cases	
Assessment	Contextualized FAIR	Data definition	FAIR maturity full review	Detailed assessments
Result	Recommendations (Objectives and Priorities)			

The preparation phase gathered the available information able to define the perimeter of the FAIR review. The landscape study provided definitions, firstly about the FAIR principles themselves, secondly about related notions like Open Science or Linked Open Data. It also retrieved information about initiatives specific to the scholarly publishing environment, like Plan S. In parallel, and closely related this time to the service provision, actual use cases allow to draw a list of potential service improvements that could be achieved by implementing FAIR principles.

The assessment phase, the most extensive one, 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, still at a rather general level. The second step listed the distinct datasets that the review will consider. The FAIR maturity full review constituted the third step in which each dataset is analyzed thoroughly according to the detailed (Wilkinson et al., 2016) as the four foundational FAIR principles are further specified through 15 recommendations[10]. 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, detailed assessments were conducted. These detailed 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 result phase produced a list of recommendations for the FAIR principles' implementation. The recommendations comprised objectives and priorities. The objectives consisted of a selection of the areas where FAIRification could be improved, whereas the priorities represent the classification of the objectives according to the service priorities in terms of feasibility, utility, and warranty. It is noteworthy that the result phase consisted of recommendations: they imply indeed other actors than the authors of the review, namely the other members of the organization and its customers.

[10] The four foundational FAIR principles are further specified through 15 recommendations. They are reported in Annexe 1 and available here: https://www.go-fair.org/fair-principles/.

These distinct phases and their corresponding steps constitute the basis for the toolkit for the FAIRification of scholarly publishing services. The following sections will detail their content.

Preparation phase: Context and Challenges

Landscape study

The landscape study described here gathered information on concepts and initiatives closely related to the FAIR principles in order to clarify definitions. The first definitional step regards 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.

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”[11] 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)[12]. At a European level, the construction of the European Open Science Cloud (EOSC)[13] strongly relies on FAIR. With the French National Plan for Open Science (Plan National pour la Science Ouverte—PNSO)[14], renewed and reinforced in 2021[15], 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.

[11] GOFAIR, “FAIR principles”: https://www.go-fair.org/fair-principles/.

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

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

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

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

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[16] 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.

Another fundamental notion related to the FAIR principles is Linked Open Data (LOD)[17]. 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[18] 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.

The last element to be considered is Plan S[19], 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).

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

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

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

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

In conclusion, the landscape study shows that the FAIRification of a publishing service takes place in a broader environment, where various options are available to better define the objectives of the FAIRification, but also where some specific constraints have to be addressed.

Use cases

The landscape study gave us some elements to establish an initial risk/benefit analysis based on actual use cases identified within the infrastructure. The risk/benefit analysis can show the first leads for the FAIR maturity review, and it is useful also at the end of the process to establish priorities. In this prospect, the OpenEdition team listed a series of actual use cases in which a systematic application of the FAIR principles would have been useful. They primarily regard the practices related to identifiers and licenses.

The first use case is the existence of parallel identifying systems. While OpenEdition's contents are available on the website, the standardized metadata is available through an OAI-PMH[20] repository. Documentary units are identified internally and in OAI, respectively, by "platform/site_name/ID" and "oai:revues.org:archeomed/7020". Therefore, modifying the name of the platform and URLs leads to an identifier modification, contrary to the principle that identifiers should be persistent. The case of 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[21]) 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 regards 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, 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 regards 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

[20] 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.

[21] 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/.

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.

Assessment phase: The FAIR principles as a grid

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 for the building of a broadly usable toolkit. For this reason, this paper may give more details about specific use cases (for example, the OpenEdition Books and Journals platforms). 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 use cases as an illustration.

Contextualized FAIR principles

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, 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 contextualisation, 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 analyse. 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 to the information system monitoring. 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[22]

Dataset	**Type**	**Schema***	**Software**	**Access****	**Creator**	**Finality**
Journals	Journals Articles Others	TEI ----- DC METS MARC ONIX	Lodel	HTML PDF ePub ----- OAI-PMH	Author Publisher	Dissemination
Books	Monographs	TEI	Lodel	HTML	Author	Dissemination

[22] Other datasets have been discarded from this table, either because of their small size, or because of their low level of FAIRness.

	Chapters Others	---- DC METS MARC		PDF ePub ---- OAI-PMH	Publisher	
Hypotheses	Blogs/posts	DC	Wordpress	HTML ---- OAI-PMH	Author	Dissemination
Calenda	Announcements	DC	Lodel	HTML ---- OAI-PMH	Author OpenEdition	Dissemination
Vocabulary	Terms	(Internal)	(OpenTheso)	HTML	OpenEdition	Enrichment
Training corpus	Enriched TEI	TEI		Github	OpenEdition	Enrichment
Metrics	Metrics		Matomo	HTML	Matomo OpenEdition	Monitoring
Catalogs	Detailed listings (books, journals, blogs)	Kbart		HTML CSV TXT XLS	Publisher 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 Annexe 2 for details). For journals, books and events, OpenEdition uses a home-built CMS, Lodel[23]. The blogs are created with the CMS Wordpress. A full-text TEI version is available for journals and books. 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 maturity review

At the core of the FAIR assessment process lies the full FAIR maturity review of each dataset. Such analytical work is necessary to avoid a generic application of the FAIR

[23] https://lodel.org/666

principles. In fact, it helps identify the actions required towards FAIR. For each dataset, the analytical table contains a short description (creator of the data, expressions of the data and of the metadata). The table also displays, for each FAIR principle, the current FAIRness status of the dataset. It also shows the existent elements that allow the FAIRification and the ones that are still lacking. A selection of these tables is reported in Annexes 2 and 3. 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 use case:

- OpenEdition Journals:

The data types include articles, issues, and collections, among others such as reviews). 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.

- OpenEdition Books:

The data types include books, chapters, and collections, but also other typesas bibliographies. 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. 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 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). It lacks at the moment the qualities to be considered a FAIR vocabulary. Nevertheless, its integration into a thesauri management tool (OpenTheso[24]) 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:

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[25]). 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.

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.

Detailed assessments

The full maturity review also highlighted areas that are crucial for publishing systems in general, and for which the OpenEdition team needed clarifications. Therefore, to ensure a complete FAIR implementation, specific assessments of some target elements were conducted. In these detailed assessments we relied on both a technical state of the art and an evaluation of the current situation of the elements at OpenEdition. The elements were the following: persistent identifiers, licensing, and author's information management.

Persistent Identifiers

As we have mentioned, the first step to ensure findability in the terms of the FAIR principles is to identify each digital object through a specific system: the persistent identifier, or PID. Like other identifiers, a PID is a non-semantic string of characters identifying a single object, but adapted to the digital environment: it is 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. The existing PID systems specifically address the persistent identification of digital objects throughout the various changes of URLs and data locations. 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. PIDs, therefore, are actually part of PID systems that are usually managed by global agencies, often for a fee. There are technically no obstacles for a

[24] 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/.

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

local organization to maintain its own PID system, but the organization's limited perimeter and/or sustainability would lower its authoritative quality.

Based on this summary about the PIDs, a few remarks can already be made when considering the publishing services. 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[26]. 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.

OpenEdition's assessment

OpenEdition has implemented the Crossref DOIs[27] for some of its contents. These PIDs imply, however, some limitations. They cannot be 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 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.

[26] One example is the deliverable "Persistent and Unique Identifiers" by CLARIN (Wittemburg, 2009).
[27] https://www.crossref.org/

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[28], ARK[29], PURL[30], and the DOIs[31] of distinct registration agencies[32]. 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[33] provides DOIs similar to Crossref DOIs for a minor cost and with a metadata schema that fits OpenEdition's needs.

Licensing

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[34], 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, France in our case.

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*[35]) 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[36] and the Open Database Licence[37]. 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

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

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

[30] https://sites.google.com/site/persistenturls/

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

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

[33] https://datacite.org/

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

[35] 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.

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

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

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 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 created by OpenEdition 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[38].

OpenEdition's assessment

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 license. However, there is no general policy for CC licensing, which can differ within a platform or from one platform 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[39] 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

[38] 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.

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

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[40] 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[41]. 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

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[42], 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[43] is a platform of the French Bibliographical Agency for Higher Education, the ABES[44]. 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[45] 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).

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

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

[42] https://info.orcid.org/what-is-orcid/.

[43] https://www.idref.fr/

[44] Agence bibliographique de l'enseignement supérieur (ABES): https://abes.fr/.

[45] Virtual International Authority File (VIAF): http://viaf.org/.

OpenEdition’s assessment

Concerning such registries, the author's information management in OpenEdition appears to be less consistent. No 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.

The detailed 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.

Result phase: Recommendations

The result 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).

Priorities

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.

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[46] 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 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.

Extended objectives

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 notably increased 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

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

This recommendation 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.

Elements for a FAIR publishing toolkit

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 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 customers. 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.

In that prospect, we can notice that, beyond the main recommendations focusing on Findability and Reusability, some aspects of the FAIR principles could be further improved in the context of OpenEdition, especially regarding machine-actionability and community standards. A solution for automated management of Accessibility through the use of an AAI still needs a better definition and planning. The Interoperability, ensured through standards in use in the publishing and library communities, could be enhanced with a better connection to the scientific community standards, such as disciplinary controlled vocabularies. 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. The provisional toolkit for the FAIRification of publishing services here presented should now itself enter into a process. It should indeed be validated, improved, and enriched by the community in order to make sure it fully addresses the specific needs of the publishing services in terms of FAIRification.

References

FitSM. 2016. “FitSM Standard for IT Service Management-Part 0: Overview and Vocabulary”.

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.

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

Annexe 1: Detailed FAIR principles

Code	**Principle**
F1	(Meta)data are assigned a globally unique and persistent identifier
F2	Data are described with rich metadata (defined by R1 below)
F3	Metadata clearly and explicitly include the identifier of the data they describe
F4	(Meta)data are registered or indexed in a searchable resource
A1	(Meta)data are retrievable by their identifier using a standardized communications protocol
A1.1	The protocol is open, free, and universally implementable
A1.2	The protocol allows for an authentication and authorisation procedure, where necessary
A2	Metadata are accessible, even when the data are no longer available
I1	(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2	(Meta)data use vocabularies that follow FAIR principles
I3	(Meta)data include qualified references to other (meta)data
R1	(Meta)data are richly described with a plurality of accurate and relevant attributes
R1.1	(Meta)data are released with a clear and accessible data usage license
R1.2	(Meta)data are associated with detailed provenance
R1.3	(Meta)data meet domain-relevant community standards

Annexe 2: FAIR analytical review example: OpenEdition Journals

FAIR review of OpenEdition journal data			
Data summary			
Data sources	Data produced through Lodel by publishers and users Can be updated (not fully controlled by the organization) Documentary units' distinct levels: text, issue and collection levels		
Data expressions	Raw data: Lodel database (as used for the HTML expression) Other expressions: TEI OpenEdition, PDF, ePUB Metadata		
Commentary	Different properties depending on the type (proper to Lodel software): - Volume contains: Publications (issues, columns, annual columns); Documentary unit contains texts, - Different types of texts (article, column, editorial, review, ...), - Annexe files types can contain data (xls, csv, sound, image, video files), Not all the different types are available in all the different expressions (TEI, pdf, epub) Question: Should the review consider the types that don't correspond to specific content (subpart, section, site, directory, etc.)?		
	FAIR implementations	**FAIR enabling information**	**FAIR limitations**
Findable			
F1. (Meta)data are assigned a globally unique and persistent identifier	Objects: - DOI (prefix 10.4000): available only for some data (depending on the types and publishers' wishes) - Issue for resources with multiple DOIs - Handles generated by Isidore harvesting platform (not retrieved by OE) Persons: a few Orcid	- OAI identifiers exist for all documentary units but are not PIDs - All documentary units are identified in the information system though the concatenation: Platform+SiteName+ID	Objects: - Some data without any PID - DOIs may exist for data published on another platform that we do not retrieve. - Handles assigned by Isidore are not retrieved. Persons: Contributors are not linked to registries.

	Organizations: a few IDs from Crossref Funding registry.		
F2. Data are described with rich metadata (defined by R1 below)	- Metadata available in the OAI-PMH repository (could be richer) - Formats DublinCore, DublinCoreTerms, METS	Rich metadata is available; could be extensively integrated in the OAI repository. 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: - ID OAI - 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): - only a selection of data is available (some types are excluded), - metadata are not complete. (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)	
Accessible			
A1. (Meta)data are retrievable by their identifier using a standardised communications protocol	HTML: accessible via the DOI Metadata: accessible via the OAI identifier		Some data without any PID (see F1)
A1.1 The protocol is open, free, and	HTTP for the data OAI-PMH for the metadata		

universally implementable			
A1.2 The protocol allows for an authentication and authorisation procedure, where necessary	All protocols are open, but not all allow for authentication. Protocol used for restricted access contents: - TCP/IP for contents requiring authentication (TEI version's case) Other protocols used where authentication is not required: - HTTP for open access contents - OAI-PMH for the metadata		Lack of a tool dedicated to the management of authentication and authorization processes.
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 references to other (meta)data	In OAI repository: - is part of - relation with OpenAIRE accessright field Some links with translations	- Citation and Cited-by available but not disseminated - Link with translations not recorded in the OAI repository - 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. 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”) The license applied to the documents is not always clear. 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 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.		

	I2: Fewer (meta)data meet disciplinary communities standards.		

Annexe 3: 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 SSH focused 188 entries covering topics, geographic areas, and periods of time. Aligned with broad categories from: CAIRN, Érudit, HAL.		
Data expressions	Used for all Calenda platform's contents. Partially used by other platforms and services. Facet of the search interface. 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 and the terms in the various languages.		Creation of descriptions for each entry, similar to Clarivate's "Scope Notes".
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	Possibility on the interface to search by the	Terms will be searchable via OpenTheso.	

searchable resource	"themes" corresponding to the vocabulary entries.		
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

Acronyms and Abbreviations

CC	Creative Commons	https://creativecommons.org/
CMS	Content Management System	https://en.wikipedia.org/wiki/Content_management_system
DC	Dublin Core Metadata Element Set	https://dublincore.org/
DOAJ	Directory of Open Access Journals	https://doaj.org/
DOI	Digital Object Identifier	https://www.doi.org/
FAIR Principles	Findable, Accessible, Interoperable, Reusable	https://www.go-fair.org/fair-principles/
FitSM	family of standards for lightweight IT service management	https://en.wikipedia.org/wiki/FitSM
HTTP	Hypertext Transfer Protocol	https://fr.wikipedia.org/wiki/Hypertext_Transfer_Protocol
ISBN	International Standard Book Number	https://www.isbn-international.org/
ISSN	Internation Standard Serial Number	https://www.issn.org/
JATS	Journal Article Tag Suite	https://jats.nlm.nih.gov/
KBART	Knowledge Bases and Related Tools	http://www.niso.org/standards-committees/kbart
LOD	Linekd Open Data	https://www.w3.org/egov/wiki/Linked_Open_Data
MARC	Machine-Readable Cataloging	https://www.loc.gov/marc/marcdocz.html
METS	Metadata Encoding and Transmission Standard	http://www.loc.gov/standards/mets/
OAI	Open Archives Initiative	https://www.openarchives.org/
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting	https://www.openarchives.org/pmh/
ONIX	Online Information exchange	https://www.ifla.org/best-practice-for-national-bibliographic

		-agencies-in-a-digital-age/node/8859
ORCID	Open Researcher and Contributor ID	https://orcid.org/
PID	Persistent Identifier	https://www.openaire.eu/what-is-a-persistent-identifier
PNSO	Plan National pour la Science Ouverte (National Plan for Open Science)	https://www.ouvrirlascience.fr/national-plan-for-open-science-4th-july-2018
RDA	Research Data Alliance	https://www.rd-alliance.org/
RDF	Resource Description Framework	https://www.w3.org/RDF/
SKOS	Simple Knowledge Organization System	https://www.w3.org/2004/02/skos/
Solr	Open source search platform	https://lucene.apache.org/solr/
TCP/IP	Transmission Control Protocol / Inernet Protocol	https://en.wikipedia.org/wiki/Internet_protocol_suite
TEI	Text Encoding Initiative	https://tei-c.org/
UNIMARC	Variation of MARC	https://www.ifla.org/publications/unimarc-formats-and-related-documentation
URI	Uniform Resource Identifier	https://fr.wikipedia.org/wiki/Uniform_Resource_Identifier
URL	Uniform Resource Locator	https://fr.wikipedia.org/wiki/Uniform_Resource_Locator
URN	Uniform Resource Name	https://fr.wikipedia.org/wiki/Uniform_Resource_Name

Bibliography

Behnke, Claudia, Luiz Bonino, Gerard Coen, Yann Le Franc, Jessica Parland-von Essen, Leah Riungu-Kalliosaari, and Christine Staiger. 2020. "D2.3 Set of FAIR Data Repositories Features," January. https://doi.org/10.5281/zenodo.3631528.

Devaraju, Anusuriya, Mustapha Mokrane, Linas Cepinskas, Robert Huber, Patricia Herterich, Jerry de Vries, Vesa Akerman, Joy Davidson, Hervé L'Hour, and Michael Diepenbroek. 2020. "M4.9 Report on Fair Data Assessment Mechanisms to Develop Pragmatic Concepts for Fairness Evaluation at the Dataset Level," August. https://doi.org/10.5281/zenodo.4118405.

Directorate-General for Research and Innovation (European Commission), and EOSC Executive Board. 2020. *Six Recommendations for Implementation of FAIR Practice by the FAIR in Practice Task Force of the European Open Science Cloud FAIR Working Group*. LU: Publications Office of the European Union. https://data.europa.eu/doi/10.2777/986252.

Directorate-General for Research and Innovation (European Commission), EOSC Executive Board, Ana Slavec, Sarah Jones, Jan Magnus Aronsen, Pedro Principe, Natalie Harrower, Françoise Genova, Oya Beyan, and András Holl. 2021. *Recommendations on Certifying Services Required to Enable FAIR within EOSC*. LU: Publications Office of the European Union. https://data.europa.eu/doi/10.2777/127253.

Directorate-General for Research and Innovation (European Commission), EOSC Executive Board, Mrio Valle, André Heughebaert, Rachael Kotarski, Tobias Weigel, Raphael Ritz, et al. 2020. *A Persistent Identifier (PID) Policy for the European Open Science Cloud (EOSC)*. LU: Publications Office of the European Union. https://data.europa.eu/doi/10.2777/926037.

Grootveld, Marjan, Joy Davidson, Angus Whyte, and René Van Horik. 2020. "D3.5 Description of FAIRsFAIR's Transition Support Programme for Repositories," September. https://doi.org/10.5281/zenodo.4058340.

Jones, Sarah, and Marjan Grootveld. 2017. "How FAIR Are Your Data?" November 24. https://doi.org/10.5281/zenodo.1065991.

Juty, Nick, Sarala M. Wimalaratne, Stian Soiland-Reyes, John Kunze, Carole A. Goble, and Tim Clark. 2020. "Unique, Persistent, Resolvable: Identifiers as the Foundation of FAIR." *Data Intelligence* 2 (1–2): 30–39. https://doi.org/10.1162/dint_a_00025.

Molloy, Laura, Josefine Nordling, Marjan Grootveld, René van Horik, Angus Whyte, Joy Davidson, Patricia Herterich, et al. 2020. "D3.4 Recommendations on Practice to Support FAIR Data Principles," June. https://doi.org/10.5281/zenodo.3924132.

Whyte, Angus, Claudia Engelhart, Daniel Bangert, Gabin Kayumbi-Kabeya, Simon Lambert, Mark Thorley, Ryan O'Connor, Patricia Herterich, and Joy Davidson. 2019. "D3.2 FAIR Data Practice Analysis," December. https://doi.org/10.5281/zenodo.3581353.

Wilkinson, Mark D., Michel Dumontier, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Mario Prieto, Dominique Batista, Peter McQuilton, et al. 2019. "Evaluating FAIR Maturity through a Scalable, Automated, Community-Governed Framework." *Scientific Data* 6 (1): 174. https://doi.org/10.1038/s41597-019-0184-5.

Wu, Mingfang, Fotis Psomopoulos, Siri Jodha Khalsa, and Anita de Waard. 2019. "Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories." *Data Science Journal* 18 (1): 3. https://doi.org/10.5334/dsj-2019-003.