Towards the Interoperability of Scholarly Repository Registries

No Author Given

No Institute Given

Abstract

Purpose – The enactment of Open Science relies on scholarly repositories capable of making research products findable and accessible. Scholarly repository registries keep authoritative metadata descriptions and persistent identifiers (PIDs) to support researchers and infrastructure providers in discovering the repositories they need and make repository managers visible to such users. However, the proliferation of repositories targeting different research products (e.g., publications, data, and software) or serving specific disciplines led to creation of many registries whose scope is not mutually exclusive. On the one hand, such a fragmented landscape forces repository managers to manually register the same repository onto multiple registries to maximise its visibility. On the other hand, researchers and scholarly infrastructure providers must elect a registry of preference or, even worse, juggle different overlapping ones whose content might diverge. While favouring a plurality of registries, this paper claims their interoperability is essential to remove the aforementioned barriers and foster their full, unambiguous exploitation.
Design/methodology/approach – We analyse the data models of four prominent registries - FAIRsharing, re3data, OpenDOAR, and ROAR - and classify their properties and overlap.
Findings – We provide a crosswalk between their data models and suggest a common data model shared across the examined registries to pave the way toward interoperability. As a means of validation, we include a coverage evaluation of the proposed data model.
Originality – To the best of our knowledge, this is the first attempt to address interoperability for scholarly repository registries.

Keywords: Scholarly Registries · Repositories · Interoperability · Open Science · Scholarly communication.

Paper type: Research paper

1 Introduction

Open Science practice heavily relies on scholarly repositories as they are central to enabling access to research outputs and granting their long-term preservation. Moreover, they are crucial for improving the visibility, discoverability and reuse of research products (Pampel et al., 2013; Wallis et al., 2013; Davidson et al., 2014; Pasquetto et al., 2019; Silvello, 2018; Bardi et al., 2022). Different types of research repositories exist, depending on the kind of content they host, the subject area or discipline they refer to, and who manages them, be they institutional repositories, data, software, or literature repositories, discipline-specific, or catch-all repositories (e.g., Zenodo and Figshare).

As the number of repositories has been remarkably increasing in the last decade, there is a growing need for services capable of providing repositories with an *identity*, possibly via persistent identifiers (i.e., PIDs), to enable non-ambiguous referencing and support tracking of provenance and impact. To this end, specialised *scholarly repository registries* (hereafter, registries for short) have been set up (Ball et al., 2014; Pampel et al., 2013; Wallis et al., 2013; Davidson et al., 2014) in order to hold authoritative *profiles* describing a broad range of information about registered repositories, to facilitate the discoverability of the most suitable one where to find or deposit the research products of interest. Some registries list repositories with a specific type of content or access condition (e.g., registries of data repositories, Open Access repositories); others list repositories with other contextual entities (e.g., databases, data policies and standards). In contrast, others list repositories regardless of the type of content (e.g., platform-specific registries).

From the repository manager's point of view, since multiple registries exist and serve different scientific domains, communities, and use cases, registering the same repository in various registries is legit, if not favourable. In fact, for the sake of visibility, repository managers are incentivised to register their repository in more than one registry at the expense of information maintainability and up-to-dateness. From the point of view of registry consumers, such a plurality of scholarly repository registries provides a full-spectrum perspective across scientific disciplines and research applications, which becomes a rich asset for researchers, scholarly service providers, and infrastructures.

However, the availability of multiple registries, repository identifiers, and profiles poses non-trivial questions and challenges regarding authoritativeness, disambiguation, and coverage. Similarly, subsequent repository registrations can show different details based on how different registries model the information required to register a repository and may reflect the bias and jargon adopted in the scientific discipline of reference. Furthermore, this fragmentation becomes a drawback in terms of information redundancy and scattering, and may lead to potential information inconsistencies across registries. In particular, this may entail non-trivial complications for provenance and impact tracking, as multiple registrations produce different identifiers for the same repository, which are then arbitrarily used across the global scholarly communication record.

In summary, from a registry user perspective, the following challenges arise:

- For repository managers, keeping different profiles manually up-to-date is a tedious task entailing filling in data entry forms adhering to different data models, a practice often leading to information redundancy, misalignment or imprecision.
- Repository registry consumers, such as researchers or scholarly communication infrastructures that need to rely on repository PIDs (e.g., the European Open Science Cloud (EOSC)[1], OpenAIRE[2], Data Management Plan services, research impact services) must cope with repository profile duplication, multiple identifiers, and heterogeneous coverage policies.

Ideally, both the challenges mentioned above can be targeted by *(i)* ensuring full interoperability among registries, which is first and foremost hampered by the diverse ways a registry can model its information, and *(ii)* by ensuring that registries can interlink their repository profiles via reciprocal identifier referencing.

In an attempt to shed some light on such an opportunity, the present paper analyses four prominent research registries – FAIRsharing, re3data, OpenDOAR, and ROAR – and focuses on a comparative analysis of their respective data models and profile overlap in order to isolate the core adherences and the substantial differences among them.

The contribution of this paper is twofold. In the first place, we provide a qualitative analysis of the four data models, documenting the information overlap and observing practical examples of repositories registered multiple times across the registries. Secondly, we deliver a crosswalk for frequently recurring informative macro-areas and suggest a minimal common data model to achieve registry interoperability.

2 Related work

The many scholarly communication services and academic search systems spun in the last decade have so far grown in parallel with a different, often siloed, mindset to target a broad and diverse range of use cases and application contexts (Bardi et al., 2022; Martín-Martín et al., 2020; Gusenbauer and Haddaway, 2020; Visser et al., 2021; Harzing, 2019). Despite modelling the very same aspects of the academic domain, the respective data models can be quite distant in order to capture the different peculiarities at hand.

Nowadays, talking about their comparison and interoperability, at least among the open alternatives, is of paramount importance to ensure the exchange of data, reuse of services, and, more recently, ensure open science practices (Burton et al., 2017; Warner et al., 2007; Aryani et al., 2020; Houssos et al., 2014; Lezcano et al., 2013; Jeffery et al., 2014).

In this work, we address a topic so far left untouched: interoperability across scholarly repository registries, which are the focal point of the present study.

[1] European Open Science Cloud – `https://eosc-portal.eu`
[2] OpenAIRE project – `https://www.openaire.eu`

The Confederation of Open Access Repositories (COAR) placed interoperability "across repositories, repositories networks and other systems and platforms" in its 2022 - 2024 strategic direction [3]. Interoperability is also one of the objectives that the International Repository Directory (IRD; *de facto*, a registry) Task Group will monitor during its activity[4], and is a key guiding principle towards Next-Generation Repositories (Rodrigues et al., 2017) which will "adopt common behaviours, functionalities and standards ensuring interoperability across institutions and enabling them to engage in a common way with external service providers".

To date, the only tangible trace towards interoperability across repository registries that we could track can be found in ROAR at the end of the registration process, where the user is asked whether the repository about to be submitted should be contextually registered on OpenDOAR.

Our work shares the greater call for interoperability in scholarly communication and the use of common data models to improve the efficiency of the different tools, the dissemination of research products and services and their possibilities of reuse and composability, as suggested in other works (Burton et al., 2017; Thomas Habing et al., 2009; Warner et al., 2007). Several efforts to achieve interoperability between registries in other domains have been recorded, for example, in the healthcare field and eGovernment.

To our knowledge, no prior study advocates for the interoperability of scholarly repository registries.

3 Data

For this study, we selected four prominent scholarly repository registries for comparison: FAIRSharing, re3data, OpenDOAR, and ROAR. In this section, we provide *(i)* the generic description of each registry; *(ii)* the methodology adopted to retrieve the data; *(iii)* the licence under which the data is redistributed; *(iv)* the references to the documentation and schemes.

It is worth noticing that not a single registry has adopted a standardised definition of what constitutes "data" as well as what it means to be a "database"; this is an ongoing question that has resulted in each registry described below adopting its own scope and requirements (Borgman, 2015). Finally, the four registries described below list resources from all research areas. For the benefit of Open Science practices, we redistribute the content of the four registries used in this work[5].

[3] COAR Strategy 2022 - 2024 – `chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.coar-repositories.org/files/COAR-Strategy-2022-and-Work-Plan-2022-.pdf`

[4] COAR Recommendations for Operations, Funding and Governance of an International Repository Directory`https://www.coar-repositories.org/files/International-Repository-Directory-Recommendations-March-2020-1.pdf`

[5] Gitea repository (this will be deposited on Zenodo when accepted) – `https://code-repo.d4science.org/miriam.baglioni/RegistriesOverlap`

FAIRsharing. Hosted at the University of Oxford in the UK and launched in 2011, FAIRsharing (Sansone et al., 2019) is a web-based, searchable portal of three interlinked registries, containing both in-house and crowdsourced, manually curated descriptions of standards, databases (including repositories and knowledge bases) and data policies, which are persistently identifiable via DOIs. FAIRsharing maps the landscape of these three resources, monitoring their relationships, development, evolution and integration; the implementation and use of standards in databases; and their adoption in data policies by funders, journals and other organisations. FAIRsharing is also an endorsed output of the RDA FAIRsharing WG[6]. FAIRsharing management combines a community-driven approach where the internal curators are supported by the maintainers of the resources themselves, which get credited via their ORCiD. The data have been downloaded using the FAIRsharing APIs[7]. As of February 2023, FAIRsharing has over 3,800 resources, of which 1,987 are databases. In this analysis, we only consider FAIRsharing's database registry, although FAIRsharing also contains standards and policy registries.

The analysis is based on the database schema[8] as provided via the read/write FAIRsharing API, and exploring the metadata fields returned via the APIs. The content, licenced under a CC-BY-SA licence[9], can be browsed by (among other fields) registry type, discipline and country, and a live statistics page provides several at-glance-views of the landscape[10].

re3data. Gone live in 2012, re3data[11] (Pampel et al., 2013) is a global registry of research data repositories from all academic disciplines, funded by the German Research Foundation[12] (DFG). The urgency of avoiding duplication of effort and serving the research community with a single, sustainable registry is mentioned on the "About" web page of the website registry, referring to the merge between re3data and Databib in 2013. Under the project "re3data.org – Community Driven Open Reference for Research Data Repositories (COREF)", the registry is going to provide "customisable and extendable core repository descriptions that are persistently identifiable and can be referred to and cited properly. This includes unique identification in machine-to-machine communication"[13]. re3data was created to meet the need for a resource dedicated specifically to data repository (Pampel et al., 2013). The existing OpenDOAR and ROAR registries mainly focused on repositories for scholarly publications and only hosted a residual share of data repositories. Furthermore, there was a need

[6] RDA FAIRsharing WG – `https://www.rd-alliance.org/group/fairsharing-registry-connecting-data-policies-standards-databases.html`
[7] FAIRsharing API documentation – `https://fairsharing.org/API_doc`
[8] FAIRsharing schema – `https://fairsharing.github.io/JSONschema-documenter/?schema_url=https://api.fairsharing.org/model/database_schema.json`
[9] FAIRsharing licence – `https://fairsharing.org/licence`
[10] FAIRsharing stats – `https://fairsharing.org/summary-statistics`
[11] re3data registry – `https://re3data.org`
[12] German Research Foundation (DFG) – `https://dfg.de`
[13] re3data mission statement – `https://www.re3data.org/about`

for a more detailed description of each research data repository, e.g., containing precise information on access and reuse conditions. The registry can be browsed by content type, discipline or country.

We retrieved the data via re3data APIs[14] The collected data follow the re3data schema version 2.2[15]. As of February 2023, it contains 3,085 data repository profiles; the registry content is released under a CC-BY licence.

OpenDOAR. Launched in 2005 thanks to a collaboration between the University of Nottingham and Lund University, funded by OSI, Jisc, SPARC Europe, and CURL, OpenDOAR[16] is a directory listing only Open Access repositories. The listed repositories are grouped into five types (Undetermined, Institutional, Disciplinary, Aggregating, Governmental), which can be browsed by type of content, software, countries and regions[17]. To be included in OpenDOAR, repositories must meet the inclusion criteria, and the submission has to be accepted by the curators' team. The submission request to OpenDOAR is carried out in two parts: first, the application is sent by filling in a form with basic information. If the admission criteria are met, further information is requested.

The data were downloaded directly via the OpenDOAR APIs[18] The collected data follow the schema accessible through their website[19], plus some additional information findable by inspecting the metadata of the downloaded records. As of Feb 2023, OpenDOAR lists 6,015 repositories worldwide; the registry content is redistributed under a CC-BY-NC-ND licence.

ROAR. The Registry of Open Access Repositories (ROAR) is hosted at the University of Southampton, UK, and it is funded by the Jisc[20]. As declared on the project's web page, its aim is "to promote the development of Open Access by providing timely information about the growth and status of repositories throughout the world". A registered account is needed to add a new repository, and the submission will be reviewed and eventually accepted or rejected with editorial comments[21]. The data have been downloaded directly from the site, choosing the option *Multiline CSV*. The CSV data formatting is consistent with the schema available online [22]. As of February 2023, it contains 5,542 data repository profiles. The registry's content can be browsed by country, year, repository type, institutional association, repository software[23] is redistributed under a CC-BY licence.

[14] re3data API documentation – `https://www.re3data.org/api/doc`
[15] re3data.org 2.2 schema – `https://www.re3data.org/schema/2-2`
[16] OpenDOAR registry – `https://v2.sherpa.ac.uk/opendoar`
[17] OpenDOAR advanced browsing – `https://v2.sherpa.ac.uk/cgi/search/repository/advanced`
[18] OpenDOAR API – `https://v2.sherpa.ac.uk/api`
[19] OpenDOAR schema – `https://v2.sherpa.ac.uk/api/metadata-schema.html`
[20] ROAR registry – `http://roar.eprints.org/information.html`
[21] ROAR registration – `http://roar.eprints.org/cgi/roar_register`
[22] ROAR schema – `http://roar.eprints.org/cgi/schema`
[23] ROAR stats – `http://roar.eprints.org/view`

3.1 Registries' content overlap assessment

A thorough study of the four registries and their content overlap can be found in the literature (Baglioni et al., 2023). In the paper, the authors started from the "same-as" equivalences of repositories provided in some cases by the registries themselves and reconstructed, with the aid of a clustering algorithm, a comprehensive disambiguation set of the repositories therein registered.

The analysis highlights a consistent content overlap across the registries, and even within the same registry (where multiple registrations of the same repository are present), posing solid premises for the present study advocating for registry interoperability.

4 Methodology

To analyse and compare the different data models, we gathered the documentation available (as reported in the Data section) from the public web pages of the registries or other sites (e.g., Zenodo), such as high-level documentation and formal schema definitions (e.g., XML, JSON, RDF). As a fallback measure, we inspected the actual repository profiles in case of absent or incomplete documentation.

By manually analysing and comparing the data models of the four registries, we noticed the presence of clusters of metadata fields documenting frequently recurring common aspects of repositories (referred to as informative *macro-areas* in the following), which are nonetheless expressed with a considerable structural and semantic variety. We identified six informative macro-areas in which metadata fields can be partitioned, which are briefly sketched as follows. *General information* contains the fields used to identify a repository profile within a registry and describe its generalities, such as IDs, names, descriptions, and URLs. *Dates and times* contains the fields describing the creation, update and (possibly) dismissal or any other timestamped detail about a repository. The macro-area *Organisations* comprehends metadata fields describing the organisations involved with the repository, for example, for its creation and maintenance, while *Content classification* contains the fields describing the discipline or scientific domain of reference for the data stored in the repository. The macro-area *Technical details* contains the fields related to APIs, versioning and any other detail meaningful for machine/programmatic interoperability, while *Legal aspects* contains the fields about the type of access (i.e., open, closed, restricted), licences or policies adopted by a repository.

5 Results

In this section, we describe the crosswalk of the four data models and show the coverage for each selected field. Finally, we describe the resulting common data model derived as a starting proposal towards the interoperability of scholarly registries.

5.1 Crosswalking registry data models by macro-areas

Here, we describe a detailed analysis of the different registry models' fields, resulting in a crosswalk that serves the ultimate goal of mapping repository profiles across registries. As previously mentioned, we defined six macro-areas of interest for this analysis, so as a first step, we selected the metadata about each macro-area for each registry. Then, we focused on finding the precise mapping for each field across registries, selecting only the fields found in at least two registries. It is worth noticing that, in general, different data models can exhibit significant variability in granularity when describing similar concepts, such as for the *Content classification* macro-area. The crosswalk analysis for each macro-area is provided in the *Crosswalk of registries data models* spreadsheet[24], which will be deposited on Zenodo if the paper will be accepted.

General information is where we found the most significant number of common fields among registries. Unsurprisingly, a high degree of alignment can be found for several fields in this macro-area. For example, repository names and homepage URLs are more or less aligned across multiple registrations of the same repository regardless of where the replicas are found (i.e., within or across registries). Notable exceptions to this observation include suffixes trailing the official name (e.g., a translation in another language, as in `[od:3833]`[25] and `[rr:11928]`) or trailing parts in the URL, yet resolving to the same web page (e.g., `[fs:2094]` and `[rd:r3d100011306]`).

Repository descriptions are also aligned across multiple registrations, even though the degree of freedom is high, e.g., whole sentences might be edited or added at any point. In the overlapping profiles `[fs:2926]` and `[rd:r3d100013300]`, the description is the same in content, despite the different wording. In `[fs:2223]` and `[rd:r3d100010685]`, the latter has a much more detailed description.

An example of non-coincident information in replicas within the same registry can be observed in OpenDOAR `[od:2546]` and `[od:2969]`. The two replicas of the same repository provide different information on the metadata field for content types: "Theses and Dissertations" in the former, and "Journal Articles; Bibliographic References; Conference and Workshop Papers; Theses and Dissertations; Reports and Working Papers; Books, Chapters and Sections" in the latter.

Despite differences in the wording, it was pretty easy to understand the role each field has in describing a repository. Indeed, three fields are in common between all registries (i.e., ID, name or title of the repository, and URL to the

[24] Crosswalk of registries data models – `https://code-repo.d4science.org/miriam.baglioni/RegistriesOverlap/src/branch/master/doc/Crosswalk%20of%20registries%20information%20models.xlsx`

[25] Hereafter, we refer to repository profiles by indicating repository registration identifiers prefixed according to the involved registry (i.e., fs: – od: – rd: – rr:). Each example links to the profile in the relevant registry in its current version, which might differ from the one observed at the time of writing. The metadata as we collect them are provided for transparency and validation of the reported examples.

homepage). Specifying a name acronym and specifying the repository type is possible in three registries out of four.

For *Dates and times*, all the registries provide temporal information about registered repositories. However, there are only two metadata fields shared in at least two registries: one referring to the registration date of the repository – for which we found a match in all the registries – and one indicating the last update of the repository – for which we found a match in three out of four registries.

Concerning *Organisations*, all the registries provide a section with information about the organisations involved in the creation/management/provision of the repository. Usually, organisation information tends to be aligned when the same organisation is present (e.g., `[od:4562]` and `[rr:14673]` or `[fs:2926]` and `[rd:r3d100013300]`), but often the number of organisations provided is not the same (e.g., `[fs:2094]` and `[rd:r3d100011306]` where the organisations provided by re3data completely contain those listed by FAIRsharing, or `[fs:1730]` and `[rd:r3d100012862]` where the opposite happens). Three out of four have the name of the organisation, the country and the homepage URL in common. FAIRsharing provides only the name via the API, but the registry owns other relevant information since it is visible through the portal. re3data also has information about the contact in the organisation, dates related to the involvement of the organisation in the maintenance of the repository, as well as the institution type and the responsibility type of the institution w.r.t. the repository. We mapped various metadata fields: the organisation name is present in all four registries, even if one of them is called organisation title. We also mapped the fields indicating additional names, alternative names and acronyms because they logically refer to additional ways to identify the repositories. An organisation's country can be expressed with its ISO code or extended name. The institution's URL is listed in all four registries, and latitude and longitude in only two.

For *Content classification*, we observed that the fields holding information about *subjects* are regulated across the four registries by a different structure (i.e., fields) and semantics (i.e., classification schemes of reference). Indeed, re3data models subjects using the DFG classification scheme[26], while ROAR leverages the Library of Congress Classification Outline[27] (LCCO). OpenDOAR instead organises its repositories according to a list of 29 content subjects, which are internally maintained and could be subject to change in the future. Finally, FAIRsharing provides subject, scientific domain, and taxonomy terminologies. Their subject[28] and domain[29] ontologies are publicly available, provide defini-

[26] DFG Classification of Scientific Disciplines, Research Areas, Review Boards and Subject Areas (2020-2024) – `https://www.dfg.de/download/pdf/dfg_im_profil/gremien/fachkollegien/amtsperiode_2020_2024/fachsystematik_2020-2024_en_grafik.pdf`

[27] Library of Congress Classification Outline (LCCO) – `https://www.loc.gov/catdir/cpso/lcco`

[28] FAIRsharing subject ontology – `https://github.com/FAIRsharing/subject-ontology`

[29] FAIRsharing domain ontology – `https://github.com/FAIRsharing/domain-ontology`

tions and unique identifiers, and are drawn from over 50 openly-available community terminologies. In re3data and FAIRsharing only, we also noted the presence of a field containing user-defined, free-text keywords and tags.

In this case, by inspecting overlapping registrations, we noted some discrepancies. Perhaps the most remarkable one regards the misalignment of subjects classifying the type of content of repositories, as we noted that different registrations of the same repository could exhibit substantially different subject classifications in many cases. Since the four registries model subjects with different classification schemes having a different number of concepts and levels of detail, this was expected to some extent.

However, it is still surprising for the following four reasons. Firstly, in some cases, the subject classification is provided in one registration, but it is unspecified in the others (e.g., the case of profiles `[od:1886]` and `[rr:3053]`).

Secondly, two registrations across different registries could map the same repository onto different, potentially disjoint, scientific disciplines such as in profiles `[fs:3118]` and `[rd:r3d100010454]`, or `[fs:2145]` and `[rd:r3d100010747]`.

Thirdly, in some cases, we noted that a relevant subject could not be assigned to a repository despite being available in the classification schema. This has been observed when more specific subjects replace generic ones as they are preferable for classification. For example, `[fs:2315]` lacks the subject "`life sciences`" present in `[rd:r3d100011894]` – despite being available in the FAIRsharing ontology – because FAIRsharing provides a hierarchical search mechanism that allows curators to use the most specific subjects relevant to their registration. A search for "`life science`" will return `[fs:2315]`, as "Biomedical Science" is one of its child terms. However, in other cases, it may happen that catch-all subjects are preferred to more specific ones, despite the latter being available. This is the case of `[rr:3190]`, which is given the subject "TK – Electrical Engineering" in ROAR according to LCCO, while being classified as "engineering" in OpenDOAR, despite the subject #15 "Electrical and Electronic Engineering" is available in its classification scheme (see profile `[od:1963]`).

Lastly, we also noted that different profiles of the same repository within the same registry might surprisingly exhibit different subject classifications, as in `[rd:r3d100011538]` and `[rd:r3d100010412]`, or `[rr:15036]` and `[rr:15039]`.

Similarly, we observed something akin to the two fields containing user-defined keywords. Here, we were not expecting a perfect alignment, nor semantic mapping across different ontologies, as the field is modelled as unregulated free-text. However, we observed a remarkable disagreement in the tags provided in some cases. For example, `[fs:3015]`, according to the tags, is a COVID-19-related repository, while its alternate registration `[rd:r3d100010882]` provides a seemingly different picture (e.g., crystallography, tomography, nuclear magnetic resonance, and so on).

Regarding *Technical details*, the fields in common across registries, although not all four, are related to the repository software, its version or the support of the data versioning. Three out of four registries also have a field for reporting a machine-readable access point to the data, although at different degrees

of abstraction. OpenDOAR and ROAR specify the access protocol associated with the API directly in the name of the metadata field, while re3data requires one specific fieldsfor the actual protocol. The usage of the field is driven by vocabularies, which can be protocol-specific (e.g., "OAI-PMH") or high-level (e.g., "REST"), depending on the use case. FAIRsharing does not provide this information via the API, but the registry knows the information since it is visible through the portal. At first glance, the information that would be provided by FAIRsharing is similar to the one provided by re3data. We have mapped all of them together because they logically refer to the same information. As for versioning, there is a field to specify if the repository is performing data versioning (in two registries out of four).

The macro area for which it was most difficult to outline a crosswalk is *Legal aspects*. Only re3data.org and FAIRsharing provide concise and detailed information on various legal concerns (e.g., data access and storage, reuse). Furthermore, the metadata referring to the policy – a field present in both re3data and OpenDOAR[30] – may express different contents, or the number of documents linked may vary in different records of the same repository. For example, `[rd:r3d100012285]` and `[od:3824]` provide different links even though they can be ascribed to the same organisation.

Looking at the licences under which the data are released, the only registries presenting comparable metadata fields are FAIRsharing and re3data. Generally, for the overlapping repositories in these two registries, we found the same information for licences – usually the same link or links to different pages under the same domain. In some cases, the Creative Commons framework is mentioned in addition to terms of use for the specific repository. However, there are also cases with discrepancies, e.g., in `[fs:2941]`, there are no details on licences or general terms of use, but in `[rd:r3d100013305]`, there are two data licences linked. ROAR is the more concise registry in these regards, as it only states whether the repository has an Open Access mandate. We could map the field expressing the data policy in re3data and OpenDOAR as they both refer to data. The data access condition field is present in three registries, but since in OpenDOAR this is open by default, the type of restriction is expressed only in two registries. In addition to repository access, a second level of information concerns data access. In this case, the items we mapped are `dataUploadType`, and `Dataset deposition type`.

The results of this groundwork allowed us to observe several remarkable "points of contact" between the registries and potential following shortcomings, which we considered an adequate motivation towards registry interoperability.

As the presence of a metadata field in a registry data model does not automatically entail its value is frequently initialised throughout every repository profile in a registry (i.e., it could be just empty in most of them), before building the common data model, we verified the coverage for each of the metadata fields included in the crosswalk.

[30] OpenDOAR only lists Open Access repositories; hence the legal information it provides is primarily external references to repositories policies.

5.2 Evaluation

In this section, we evaluate the crosswalk by quantitatively assessing its mapping coverage against the content of the four registries taken into analysis. The coverage is shown in table 1. The table shows only the fields with coverage greater than zero in more than two registries (e.g., we excluded the latitude and longitude field, as in OpenDOAR it is always empty).

Some of these fields are completely covered in the four registries: the repository name and URL, the internal identifier, and the organization name (almost all records contain information in these fields). In contrast, others are present and filled in all four registries but with variable degrees of coverage: subject (from 24,4% of ROAR to the more than 97% of the other three registries), organization name (from 21% of OpenDOAR to more than 82% for ROAR, with FairSharing and re3data having almost full coverage) and start date (from 85,25% of FAIRsharing to 99,7% for ROAR and full coverage for the other two). Nine fields are featured in 3 registries out of four - but, again, with some variability in the number of records in which they are compiled. In this case, the most covered attributes are the ones holding the repository type and description. The field holding the record URI is wholly covered in OpenDOAR (100%) and almost in FAIRsharing (94.6%), but much less in re3data (36.6%) while always blank in Roar. The additional name of the repository is filled in all FAIRsharing records, most of re3data records (78%; 2,436 out of 3,085), and far less covered in OpenDOAR, where 1,289 records have this information out of 6,015 (21.4%). The field retaining the last update date is not present in FAIRsharing only, while all the records of the other three registries cover this information in each repository profile. In OpenDOAR, only a few records provide policy details: out of 6,015 records, 760 records exhibit the policy URL (12.6%) and 720 the policy type (11.9%). While there is high coverage of all the different policy fields in re3data, a higher asymmetry is visible in FAIRsharing, with quite a small coverage for many of them.

An aspect worth mentioning is that, although some properties are present in the data models of the registries, their values are not exposed via the APIs (e.g., in FAIRsharing, the URL for programmatic access to the repository, or the identifier of a relevant organisation). Nonetheless, a direct (manual) inspection of repository profiles on the registry website shows that the information is indeed present and, therefore, could be mapped onto a common data model. This entails that, while we cannot evaluate the coverage for certain fields thoroughly, we can rest assured that the property could be initialised and potentially take part of the interoperability framework.

5.3 Common data model

We used the crosswalk to define a minimal common data model (see Supplement A) that all the registries should share. The suggested common data model results from the crosswalk and extends it in a few directions, as we have included some

Table 1: Metadata records coverage in the four registries. For each field included in the common data model, we reported the number of occurrences encountered for that information in the content of the four registries. As for OpenDOAR, the content is "open" by definition; we reported an asterisk (*) to indicate those cases where the fields in the common data model could be initialised by default.

Macro area	Field	fs	r3	rr	od
General information	identifier	1,987	3,085	5,542	6,015
	name	1,987	3,085	5,540	6,011
	nameLanguage	–	3,085	–	5,881
	additionalName	1,987	2,436	–	1,289
	repositoryType	–	3,066	5,542	6,015
	repositoryDescription	1,987	3,085	3,908	–
	repositoryContent	–	3,077	–	5,872
	repositoryURL	1,987	3,085	5,534	6,015
	recordURI	1,880	1,131	–	6,015
	recordCount	–	–	2,287	4,163
Dates and times	startDate	1,694	3,085	5,527	6,015
	lastUpdate	–	3,085	5,542	6,015
Organisations	organizationId	–	1,881	–	2,842
	organizationName	1,946	3,084	4,549	1,260
	organizationNameLanguage	–	3,084	–	1,249
	organizationAcronym	–	2,655	–	112
	organizationCountry	–	3,084	5,225	6,014
	organizationUrl	–	3,084	4,367	5,810
Content classification	subject	1,953	3,080	1,350	5,849
	keyword	680	3,076	–	–
Technical details	softwareName	–	2,389	4,787	5,589
	versioning	208	3,085	–	–
	apiURL	–	1,406	4,415	4,572
Legal aspects	policyURL	–	4,933	–	760
	accessType	362	3,085	–	*
	accessRestriction	124	162	–	–
	dataUploadType	278	3,068	–	*
	dataUploadRestriction	185	2,045	–	–
	licenseName	1,238	3,085	–	–
	licenceUrl	1,238	3,085	–	–

noteworthy properties that might not be present in the four registries' models – or at least not in all of them.

One essential aspect, mainly for the automatic processing of the information, is the presence of the vocabulary to which a given value refers. Hence, all the fields related to language and country code should refer to a reference vocabulary (e.g., "ISO-639-3"). The same holds for the repository subject and keywords: in our common data model, there are attributes specifying the code, the schema from which the code was derived, and the URI associated with the value.

Still, to facilitate the automatic metadata or repository content processing, we added the type of API provided by the repository as a required field. A documentation URL can also be provided.

Other metadata fields that can give important information are those about the organisations related to the registered repository. In particular, these include "organisation role" and "organisation type". The organisation's role encodes, for example, whether the organisation is responsible for maintaining, funding or contributing to the repository. The organisation type instead encodes, for example, whether it is a higher education company or a funding organisation.

It is also worth mentioning that the preferred identifier for the repository should be a PID (e.g., DOI) and not just an identifier internal to the specific registry (as in the case of ROAR and OpenDOAR), as this would not enable users always to reference the repository correctly.

Notice that all the above properties are optional in the common data model.

We decided to model the information regarding licence and access policies as generic as possible for the section legal aspect. Each of them is modelled as an element specifying the kind of information, with two attributes providing specifics for the value.

At least two registries share all the other fields in the common data model.

6 Discussion

As emerged from the overlap analysis, the four scholarly repository registries at hand are indeed affected by a problem of information redundancy and scattering, as repositories can be registered multiple times across them. Therefore, quantitative evidence supports the motivation underpinning the need for an interoperability framework in order to achieve a seamless flow of information across registries and synchronise a shared core of information across repository profiles.

To this end, we conducted a comparative analysis of the data models of the four registries. Here, a certain variety emerged across the schemes. However, while some degree of variety was expected, given that each registry was created to serve specific communities or different goals, the analysis revealed that there is also a disparity in terms of content conveyed even by supposedly similar and comparable metadata fields. This variety expresses a difference in the granularity of the information collected by the different registries – as in the case of information on legal aspects – as well as some discrepancies in the content –

for example, in the classifications of content and subjects. This aspect can undoubtedly confuse registries' users, such as researchers, or create difficulties in building new services for which arbitrary choices are often needed (e.g., when a scholarly infrastructure aggregates the content of registries).

The results highlighted by our analysis across registries' data models allowed us to define a crosswalk and a common data model for interoperability across registries. In the suggested common data model, most metadata fields are already present in at least two of the four registries we analysed. Since the purpose of this work is to encourage effective interoperability between scholarly repository registries, we have defined the properties of this common data model and how they can be expressed (formats, possible vocabularies, number of occurrences, etc.) in order to facilitate the exchange of information between registries, but also the development of sound scholarly communication services on them. In some cases, the properties we have added can foster clarity and, therefore, greater usability of the information by users at large. The resulting common data model recommendations can be found in Supplement A, which describes it at a high level. For the sake of clarity, an XML example is provided as well in Supplement B.

The proposed common data model is intended to act as a basic guideline to enable pragmatic interoperability of metadata fields that currently cannot interoperate due to different modelling assumptions and the use of different vocabulary, classifications or even a more basic lack of common terminology (e.g., the subject field).

To the best of our knowledge, the guidelines and recommendations outlined here are the first attempts in the literature to pave the way to support interoperability across scholarly repository registries and make possible their full exploitation in a plethora of different applications.

Acknowledgements

This work was partially funded by the EC H2020 projects OpenAIRE-Nexus (Grant agreement 101017452) and EOSC Future (Grant agreement 101017536). Moreover, we are deeply grateful to Dr. Allyson Lister and the FAIRsharing team for the suggestions they provided us with.

Contributionship

Andrea, Gina, and Miriam contributed to the conception of the idea, the collection of data, the analysis of the results, the design of the common data model, and the writing of the paper. Paolo seeded the idea and supervised the work.

Bibliography

Amir Aryani, Martin Fenner, Paolo Manghi, Andrea Mannocci, and Markus Stocker. Open Science Graphs Must Interoperate! In *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium*. Springer, 2020.

Miriam Baglioni, Andrea Mannocci, Gina Pavone, Michele De Bonis, and Paolo Manghi. (semi)automated disambiguation of scholarly repositories. In *Proceedings of the 19th Conference on Information and Research Science Connecting to Digital and Library Science*, 2023. URL `https://ceur-ws.org/Vol-3365/paper2.pdf`.

Alexander Ball, Kevin Ashley, Patrick McCann, Laura Molloy, and Veerle Van Den Eynden. Show me the data: The pilot UK Research Data Registry. *International Journal of Digital Curation*, 9(1):132–141, 2014. ISSN 1746-8256. https://doi.org/10.2218/ijdc.v9i1.307.

Alessia Bardi, Paolo Manghi, Andrea Mannocci, Enrico Ottonello, and Gina Pavone. A primer on open science-driven repository platforms. In *MTSR 2022 - International Conference on Metadata and Semantics Research, Londra, UK, 07-11/11/2022*, 2022.

Christine L. Borgman. *Big Data, Little Data, No Data*. The MIT Press, 2015. ISBN 978-0-262-32786-2. https://doi.org/10.7551/mitpress/9963.001.0001.

Adrian Burton, Hylke Koers, Paolo Manghi, Markus Stocker, Martin Fenner, Amir Aryani, Sandro La Bruzzo, Michael Diepenbroek, and Uwe Schindler. The scholix framework for interoperability in data-literature information exchange. *D-Lib Magazine*, 23(1-2), 2017. ISSN 10829873. https://doi.org/10.1045/january2017-burton.

Joy Davidson, Sarah Jones, and Laura Molloy. Big data: The potential role of research data management and research data registries. In *IFLA WLIC 2014*, Lyon, France, 2014.

Michael Gusenbauer and Neal R. Haddaway. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods*, 11(2):181–217, 2020. ISSN 1759-2887. https://doi.org/10.1002/jrsm.1378.

Anne Wil Harzing. Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? *Scientometrics*, 120(1):341–349, 2019. ISSN 15882861. https://doi.org/10.1007/s11192-019-03114-y.

Nikos Houssos, Brigitte Jörg, Jan Dvořák, Pedro Príncipe, Eloy Rodrigues, Paolo Manghi, and Mikael K. Elbæk. OpenAIRE Guidelines for CRIS Managers: Supporting Interoperability of Open Research Information through Established Standards. *Procedia Computer Science*, 33:33–38, January 2014. ISSN 1877-0509. https://doi.org/10.1016/j.procs.2014.06.006.

Keith Jeffery, Nikos Houssos, Brigitte Jörg, and Anne Asserson. Research Information management: The CERIF approach. *International Journal of*

Metadata, Semantics and Ontologies, 9(1):5–14, 2014. ISSN 1744263X. https://doi.org/10.1504/IJMSO.2014.059142.

Leonardo Lezcano, Brigitte Jörg, Brian Lowe, and Jon Corson-Rikert. Promoting International Interoperability of Research Information Systems: VIVO and CERIF. page 14, 2013. https://doi.org/10.3217/JUCS-019-12-1854.

Alberto Martín-Martín, Mike Thelwall, Enrique Orduna-Malea, and Emilio Delgado López-Cózar. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics*, (0123456789), September 2020. ISSN 0138-9130. https://doi.org/10.1007/s11192-020-03690-4.

Heinz Pampel, Paul Vierkant, Frank Scholze, Roland Bertelmann, Maxi Kindling, Jens Klump, Hans-Jürgen Goebelbecker, Jens Gundlach, Peter Schirmbacher, and Uwe Dierolf. Making Research Data Repositories Visible: The re3data.org Registry. *PLOS ONE*, 8(11):e78080, November 2013. ISSN 1932-6203. https://doi.org/10.1371/journal.pone.0078080.

Irene V. Pasquetto, Christine L. Borgman, and Morgan F. Wofford. Uses and Reuses of Scientific Data: The Data Creators' Advantage. *Harvard Data Science Review*, 1(2), November 2019. https://doi.org/10.1162/99608f92.fc14bf2d.

Eloy Rodrigues, Andrea Bollini, Alberto Cabezas, Donatella Castelli, Les Carr, Leslie Chan, Chuck Humphrey, Rick Johnson, Petr Knoth, Paolo Manghi, Lazarus Matizirofa, Pandelis Perakakis, Jochen Schirrwagen, Daisy Selematsela, Kathleen Shearer, Paul Walk, David Wilcox, and Kazu Yamaji. Next Generation Repositories: Behaviours and Technical Recommendations of the COAR Next Generation Repositories Working Group. November 2017. https://doi.org/10.5281/zenodo.1215014.

Susanna-Assunta Sansone, Peter McQuilton, Philippe Rocca-Serra, Alejandra Gonzalez-Beltran, Massimiliano Izzo, Allyson L. Lister, and Milo Thurston. FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology*, 37(4):358–367, April 2019. ISSN 1546-1696. https://doi.org/10.1038/s41587-019-0080-8.

Gianmaria Silvello. Theory and Practice of Data Citation. *Journal of the Association for Information Science and Technology*, 69(1):6–20, January 2018. ISSN 23301635. https://doi.org/10.1002/asi.23917.

Thomas Habing, Janet Eke, Matthew A. Cordial, William Ingram, and Robert Manaster. Developments in Digital Preservation at the University of Illinois: The Hub and Spoke Architecture for Supporting Repository Interoperability and Emerging Preservation Standards. *Library Trends*, 57(3):556–579, 2009. ISSN 1559-0682. https://doi.org/10.1353/lib.0.0052. URL `http://muse.jhu.edu/content/crossref/journals/library_trends/v057/57.3.habing.html`.

Martijn Visser, Nees Jan van Eck, and Ludo Waltman. Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. page 22, 2021. https://doi.org/10.1162/qss_a_00112.

Jillian C. Wallis, Elizabeth Rolando, and Christine L. Borgman. If We Share Data, Will Anyone Use Them? data Sharing and Reuse in the Long

Tail of Science and Technology. *PLoS ONE*, 8(7), 2013. ISSN 19326203. https://doi.org/10.1371/journal.pone.0067332.

Simeon Warner, Jeroen Bekaert, Carl Lagoze, Xiaoming Liu, Sandy Payette, and Herbert Van de Warner. Pathways: Augmenting interoperability across scholarly repositories. *International Journal on Digital Libraries*, 7(1-2):35–52, October 2007. ISSN 1432-5012, 1432-1300. https://doi.org/10.1007/s00799-007-0016-7.