CURRENT STATE AND FUTURE DIRECTIONS FOR OPEN REPOSITORIES IN EUROPE

December 2023

Kathleen Shearer, COAR | **Silvia Nakano**, COAR | **Eloy Rodrigues**, University of Minho
Natalia Manola, OpenAIRE | **Martine Pronk**, LIBER | **Vanessa Proudman**, Sparc Europe

Executive Summary

"I think repositories are absolutely necessary as part of the chain to create a sustainable Open Access world." - Survey respondent

Open Science is ushering in a new paradigm for research; one in which all researchers have unprecedented access to the full corpus of research for analysis, text and data mining, and other new research methods. A prerequisite for achieving this vision is a strong and well-functioning network of repositories that provides human and machine access to the wide range of valuable research outputs. It will require transitioning repositories from isolated institutional services towards the vision of the next generation repository, whereby repositories are part of a distributed, globally networked infrastructure for scholarly communication, on top of which layers of value-added services can be deployed.

In January 2023, OpenAIRE, LIBER, SPARC Europe, and COAR launched a joint strategy aimed at strengthening the European repository network. Through this strategy we are committed to working together - and with other relevant organisations - to develop and execute a plan that will reinforce and enhance repositories in Europe. As a first step, a survey of the European repository landscape was undertaken in February-March 2023.

The survey had **394 responses** from repositories in **34 countries**. We found that, collectively, European repositories acquire, preserve and provide open access to tens or possibly hundreds of millions of valuable research outputs and represent critical, not-for-profit infrastructure in the European open science landscape. They are used for sharing articles that may be paywalled in published journals, but also for providing access to a large variety of other types of research outputs including research data, theses/dissertations, conference papers, preprints, code, and so on.

A large proportion of repositories are based at universities making them quite sustainable and, by every indication, their collections are being well-used by the research community and beyond. The number and range of value-added services to which repositories are contributing demonstrates that European repositories have been progressing towards the vision of the next generation repository, which is about moving beyond the repository as an institutional service, to the networked repository that is an integral part of the broader ecosystem. In

addition, repositories are well placed to support the expansion of open science practices across Europe and the reformation of research assessment, which places a greater emphasis on inclusiveness, diversity, and transparency.

However, to fully achieve our vision, there is still work to be done. The survey has exposed a number of important areas where the current repository landscape could be strengthened. In particular, we found that repositories struggle **with three main challenges**:

(1) maintaining up-to-date, highly functioning software platforms,
(2) applying consistent and comprehensive good practices in terms of metadata, preservation, and usage statistics; and
(3) gaining appropriate visibility in the scholarly ecosystem.

Despite the challenges, the current climate offers exciting opportunities for repositories. Many funders are actively promoting the repository route for articles because of their role in supporting equitable access to content (i.e. no fees to access or deposit). The value proposition for open science is growing and repositories are increasingly recognised as the main mechanism for collecting and providing access to a wide range of other research outputs. Add to this, the nascent, but growing, interest in the publish-review-curate model in which repositories have a central function[1], and it seems they are well placed to expand their current role in the ecosystem.

To support this evolving role for repositories, OpenAIRE, LIBER, SPARC Europe and COAR have **identified three areas where we can work together to help advance and strengthen repositories in Europe**:

1. Highlighting the value proposition and advocating for the critical role of repositories in Europe
2. Propagating best practices for repositories across the continent
3. Assisting with the creation and coordination of national networks

In the coming months, our organisations will develop more concrete plans for advancing each of these areas.

[1] From cOAlition S: To illustrate how a scholar-led communication system can (and already does) work in practice and supports the principles of Open Science, we highlight the Publish, Review, Curate (PRC) model, which we find particularly promising. https://www.coalition-s.org/wp-content/uploads/2023/10/Towards_Responsible_Publishing_web.pdf

Table of Contents

Introduction

Open Science is ushering in a new paradigm for research; one in which all researchers have unprecedented access to the full corpus of research for analysis, text and data mining, and other new research methods. A prerequisite for achieving this vision is a **strong and well-functioning network of repositories** that provides human and machine access to the wide range of valuable research outputs. This will require transitioning repositories from isolated institutional services towards the vision of the next generation repository, whereby repositories are part of a distributed, globally networked infrastructure for scholarly communication, on top of which layers of value-added services can be deployed.

Yet, progress towards this vision has been relatively slow, and many repositories continue to struggle with older technologies and a number of other challenges. To address this COAR and other key stakeholders in different regions and countries have been working together to adopt strategies that will strengthen repository networks and accelerate the adoption of leading-edge functionalities[2].

Currently, Europe has one of the most well-developed networks globally with hundreds of repositories hosted by universities, research centres, government departments, and not-for-profit organisations. However, there are significant variations across the European repository landscape. For Europe to maintain its position as a global leader in open science, we must ensure there is a strong and sustainable network of open repositories.

In January 2023, OpenAIRE, LIBER, SPARC Europe, and COAR launched a joint strategy aimed at strengthening the European repository network. Through this strategy we are committed to working together - and with other relevant organisations - to develop and execute an action plan that will reinforce and enhance repositories in Europe.

As a first step, a survey of the European repository landscape was undertaken in February-March 2023. The aim of the survey was to gain a better understanding of the repository ecosystem in Europe. The survey was designed and disseminated by partner organisations through various channels including website announcements, email lists, twitter (X) and other social media.

This report provides the results of the survey and will assist the organisations in developing relevant and effective activities to strengthen repositories in the region.

[2] https://www.coar-repositories.org/news-updates/what-we-do/regional-initiatives/

Results

For the purposes of this survey, an open repository was defined as a digital management system that collects one or more types of research output and provides free access to the content to all users (with the exception of restrictions for sensitive data).

Number of respondents

There were **394 responses** from **34 countries** in Europe (Figure 1), with 10 countries (Austria, Croatia, Germany, Italy, Poland, Portugal, Serbia, Spain, Switzerland, UK) that each had over 15 responses. In certain areas, we provide a small snapshot of certain results of each of these countries and have undertaken a more in-depth analysis of the situation.

Figure 1: Geographic distribution of survey respondent repositories

Types of institutions

Most respondent repositories were based at universities, followed by research centres (Figure 2). The rest fell into the "other" category, which was composed of

a diversity of institution types including libraries, university departments, scientific institutions, hospitals, government entities and not-for-profit organisations. Two respondent repositories were managed by publishers. As many university repositories are managed by the library, we assume that a number of the respondents that indicated the repository was based at a university, was also located in the library.

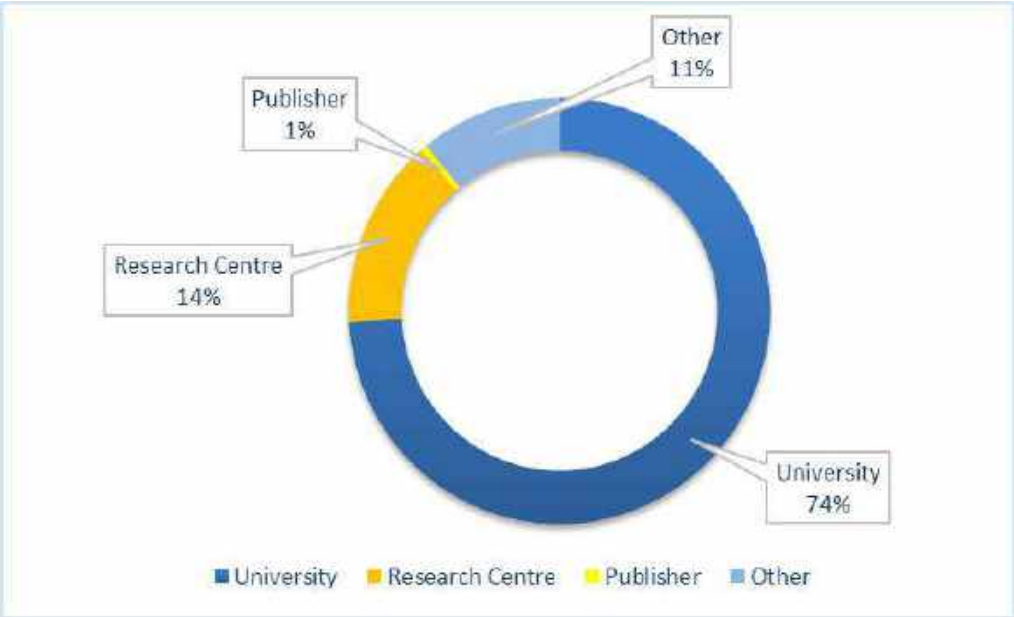

Figure 2: Types of institutions where repositories are based

Predominant content types in the repository

Most repositories reported collecting a variety of content types, with 54% of respondents indicating that the predominant content type in the repository was published articles (Table 1). Theses and dissertations are predominant for 19% of respondents and research data for 13%. 14% of respondents indicated preprints were in the top 3 of their content types, but only 1% (5 repositories) reported they were the predominant type.

Repositories with research data as their predominant content type tend not to collect publications, theses, preprints, and other - rather they seem to specialise in research data only. The repositories that collect predominantly publications (articles, theses, and preprints) usually collect a variety of content types, including research data. (Figure 3)

Respondents were not asked to specify content types if they chose the other category, so we do not have further information about what they are.

Table 1: Top three most predominant content types in repositories

	Published articles	Preprints	Research data	Theses & dissertations	Conference proceedings	Other
1st	213	5	52	74	10	20
	54%	1%	13%	19%	3%	5%
2nd	68	23	11	71	97	57
	17%	6%	3%	18%	25%	14%
3rd	14	26	19	80	89	42
	4%	7%	5%	20%	23%	11%
Top 3	75%	14%	21%	57%	50%	30%

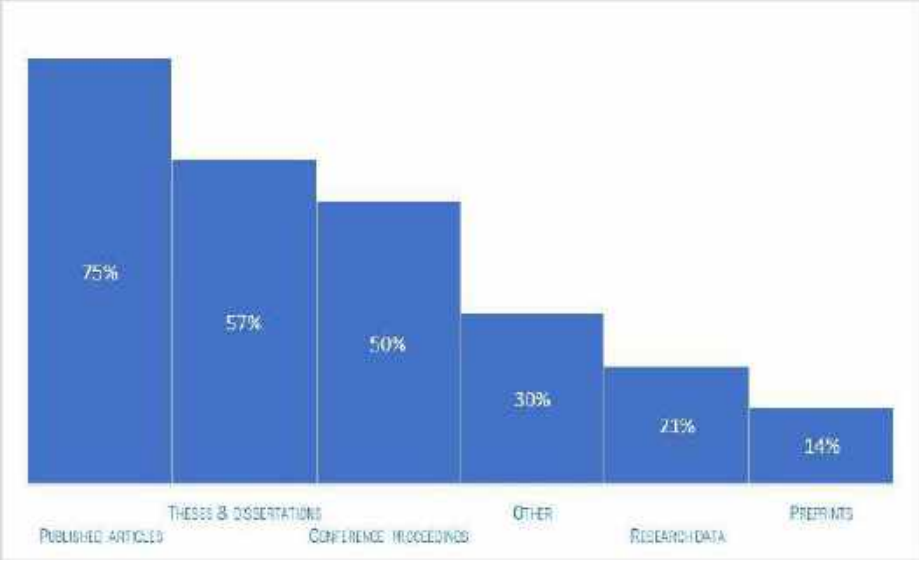

Figure 3: Top six predominant types

Number of items in the repository

Collection sizes (number of items in each repository) vary significantly across respondent repositories with about 20% having less than 1,000 items, and the six largest repositories having more than a million records each (Figure 4). The largest repository, Europe PMC, contains over 8.5 million full text records. The most frequent collection sizes of repositories are from 1,000 to 10,000 items (32.5%); 10,000 to 50,000 (27.5%); and less than 1,000 items (21.8%). The average collection size for institutional repositories is 64,859 items (repositories that collect their local research outputs), and for other repositories (domain, data, and national repositories) the average is 386,088 items.

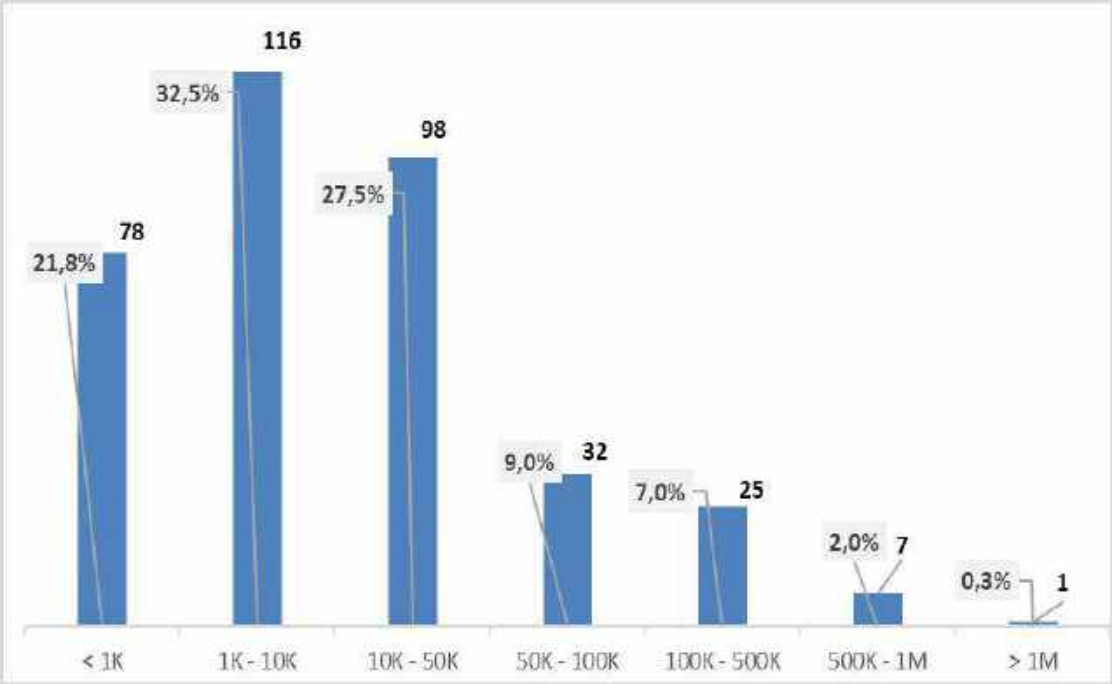

Figure 4: Repository collection sizes

Languages of metadata and content

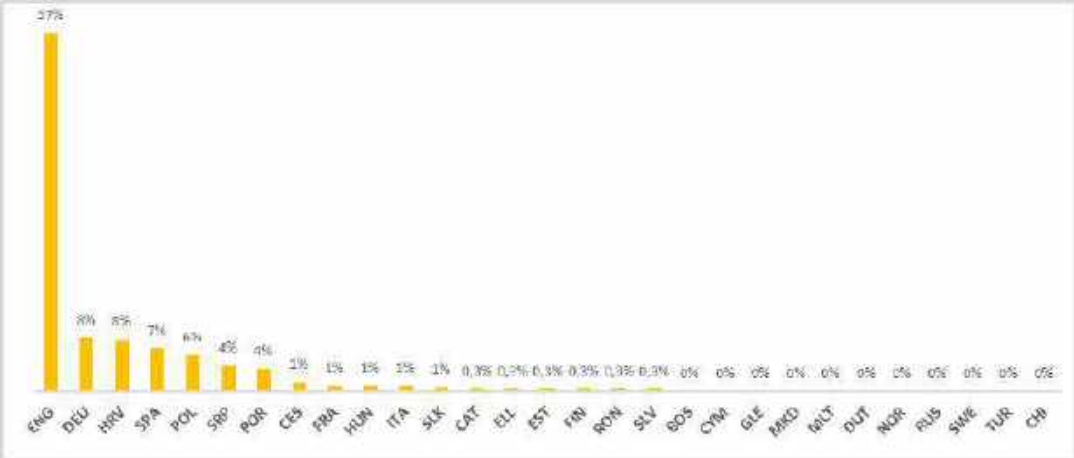

Figure 5: Predominant language of resources in repositories

For 57% of the repositories, the predominant language of the repository records is English. If we exclude UK repositories (62 respondents), 47% (142 of 299) of repositories reported that English was the predominant language of content. (Figure 5 and 6).

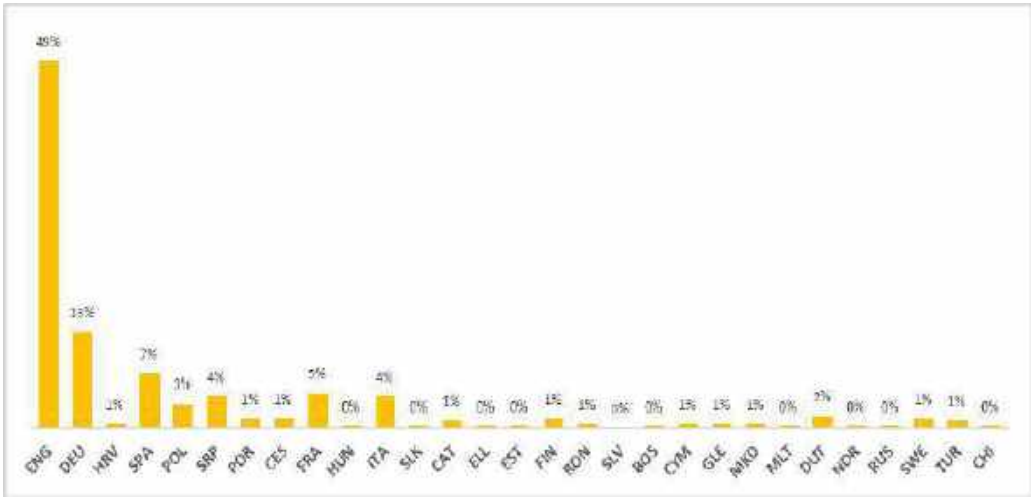

Figure 6: Second most predominant language of resources in repositories

For repositories whose predominant language is not English, it is always the national language that was reported as being predominant. For the majority of those repositories, English is the second most predominant language, with a few exceptions shown in the table below (Table 2). If we look at the countries with more than 15 repositories represented in the survey, certain ones were notable for having low portions of English content: Croatia, Portugal, Poland, and Spain.

Table 2: Languages of metadata and resources in the repositories

COUNTRIES WITH OVER 15 REPOSITORIES IN SURVEY	ENGLISH PREDOMINANT LANGUAGE		LOCAL PREDOMINANT LANGUAGE		OTHER SECOND LANGUAGES
	ENG %	% LOCAL as 2nd predominant language	LOCAL %	% ENG as 2nd predominant language	
Croatia	9%	6%	91%	69%	Italian
Portugal	22%	22%	78%	67%	Spanish
Poland	32%	26%	68%	59%	
Spain	34%	32%	66%	59%	Catalan (3)
Austria	41%	35%	59%	53%	Hungarian
Serbia	42%	42%	58%	45%	Russian
Germany	56%	42%	44%	36%	
Switzerland (*)	70%	75%	20%	15%	
Italy	80%	67%	20%	20%	Spanish
United Kingdom	100%				Spanish (6) French (5) German (3) Welsh (2) Polish, Italian & Chinese -a repository belonging to the international publisher)

(*) Local LANGUAGES: DEU, ITA, FRA

Who can deposit

Over 75% of repositories in the survey serve their local communities and offer services to only persons who are affiliated with their institution (Figure 7). 6% of respondent repositories are open to anyone, 4% are open to domain communities, and 1% are open to persons from a specific country. Most of the 9% who chose the ''other' category, clarified that the repository was an institutional repository offering a mediated deposit service, whereby repository staff deposited content on behalf of the creators, therefore the portion of institutional repositories was actually over 80%.

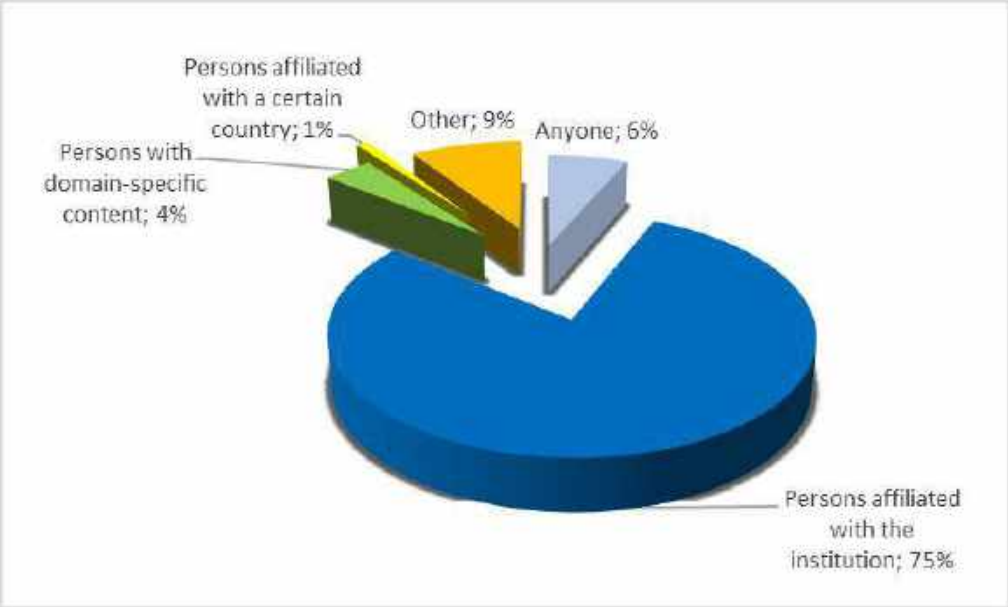

Figure 7: Repository accepts content from which communities

National networks

About half of respondent repositories indicated they were part of a national level network or service (Figure 8). The types of services/networks are varied and include harvesters, portals and other discovery/indexing services; communities of practice; shared platforms; open source platform networks; and domain networks. However, the responses were inconsistent in many countries, with some respondents from a given country indicating they belong to a network and others indicating they did not. This could be because respondents had a different interpretation of what is a national network or national services, but also some national networks may serve only a subset of repositories in their country.

However, a substantial amount, i.e. almost half of all responding repositories in a country, feel part of an existing network. Several repositories belong to more than one type of national network or service. Given the fact that the communities advancing open access and research data management communities are often distinct from each other, it is not surprising that respondents from these different sectors named different national services.

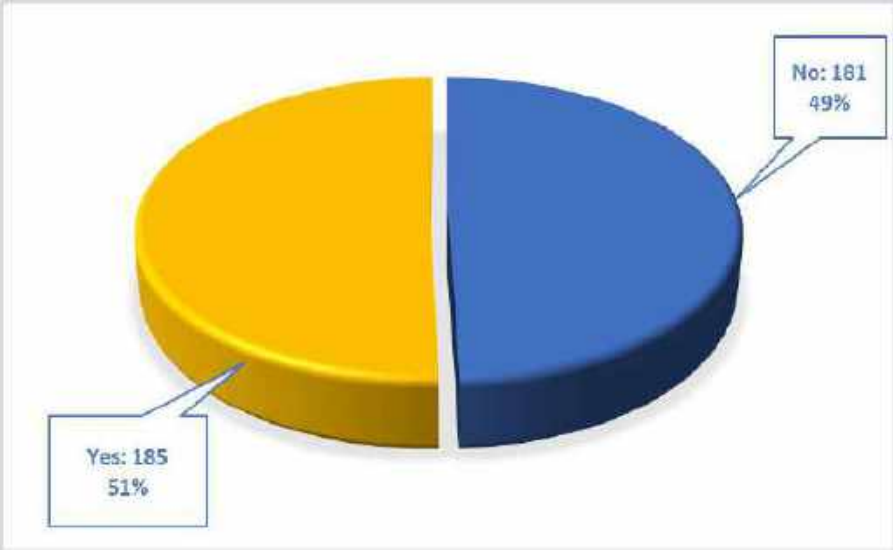

Figure 8: Number of respondents who are part of a national network

Hosting model for repository

57% (223) of respondent repositories are locally hosted, while 43% (165) of respondent repositories are hosted by an external provider (Figure 9). Most external providers are national hosting platforms, university data centres, or national cloud services. 7 respondent repositories are hosted by commercial providers.

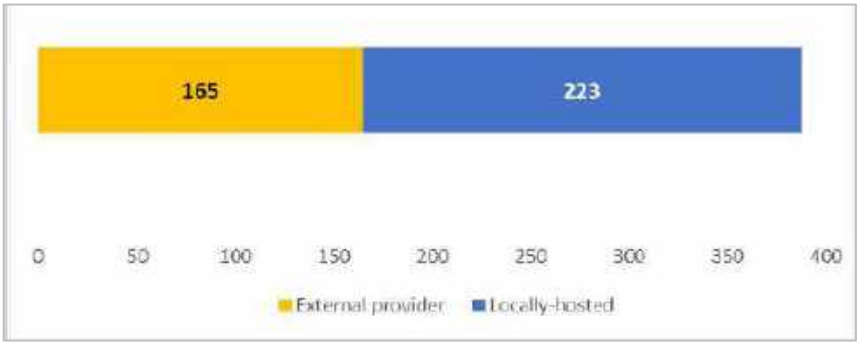

Figure 9: Local or external hosting of repository

Software platforms

DSpace is the most commonly used software platform, with 41% of respondents indicating they currently use the DSpace software. Other widely used platforms are Eprints (11%), Fedora/Islandora (11%) and Dataverse (4%). Following this, several other platforms were also reported: Invenio (3%), Pure (3%), OPUS (3%), Omega-PSIR (2%), Samvera (1%), and Figshare (1%) along with a variety of other software types. (Figure 10) It is worth noting that 8% of respondents run their repositories on locally developed software (4% of institutional repositories use a locally developed software platform and 22% of national / domain / generalist repositories have locally developed software platform).

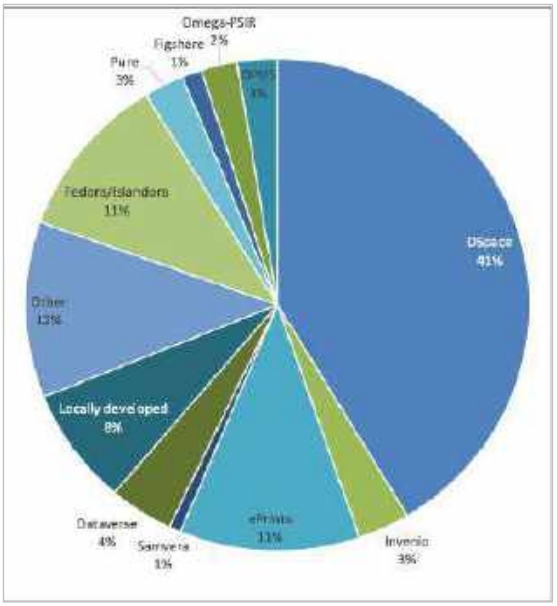

Figure 10. Software platforms used by repositories

Add-ons/patch/code added to the codebase

About 61% of all respondents indicated that they have changed or added to the basic "out of the box" versions of the repository software platform (Figure 11).

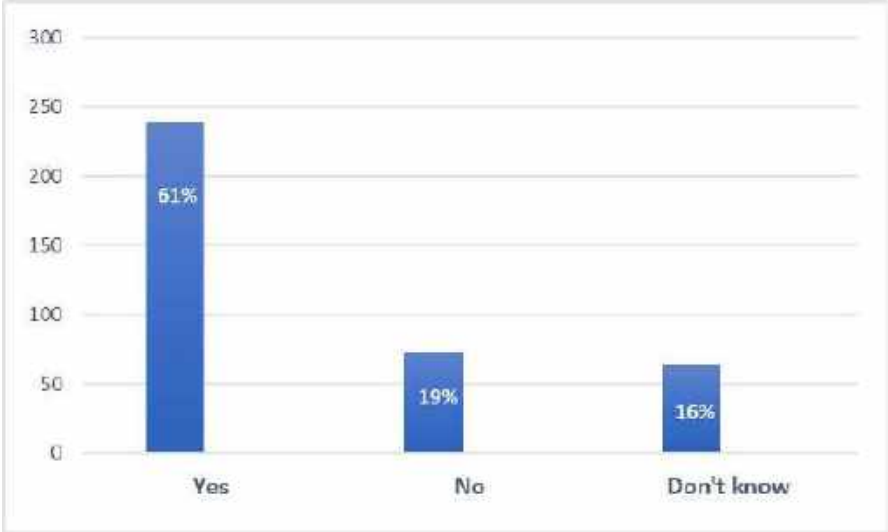

Figure 11: Number of respondents that adopt add-ons, patches of new code

This situation is more frequent in Eprints (83,7%) and DSpace repositories (63,2%), compared with all the other platforms (58%).

Software Upgrades

42% of repositories upgraded their repository platforms in 2022, and 74% of repositories stated that they were planning to upgrade in 2023. 21% of repositories that upgraded in 2022, plan to do it again in 2023. In total, about 60% of respondents have either updated their repository in 2022 or are planning to update to a more recent version in 2023 (Figures 12 and 13).

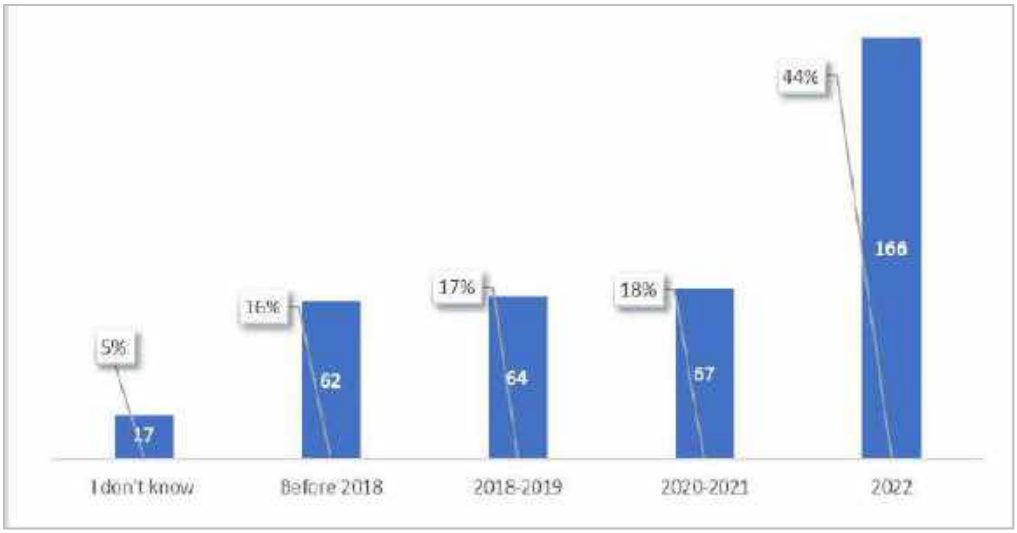

Figure 12: Year of last major upgrade

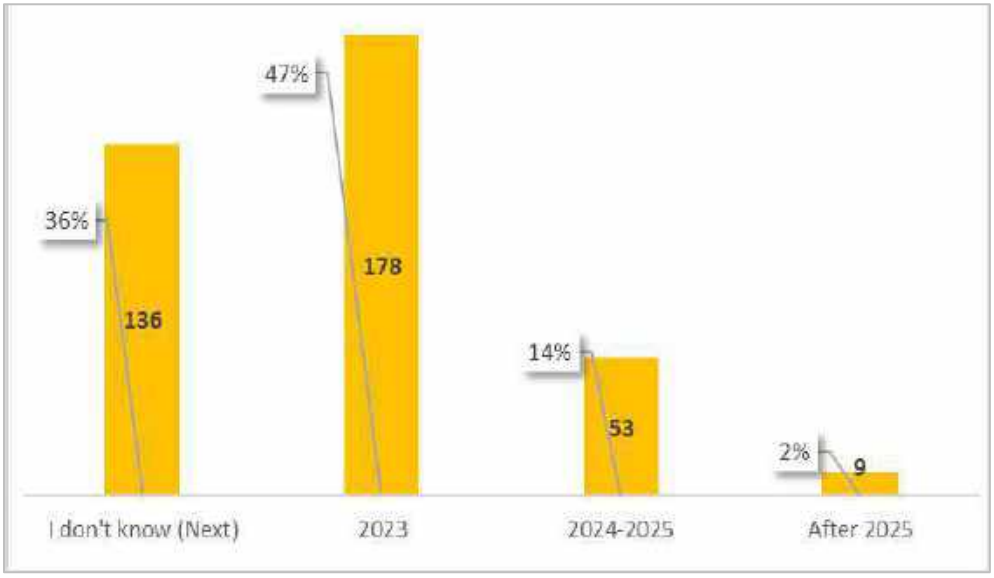

Figure 13: Year of next major upgrade

Metadata schemas

The most common metadata schema adopted in repositories is Dublin Core, with 77% of repositories indicating they provide support for Dublin Core (Figure 14). 26% provide support for the DataCite schema, which was initially developed for research data and unsurprisingly, there was a positive correlation between the repositories that collect research data and support the DataCite schema. Just under half of respondents indicated that they support more than one type of metadata schema.

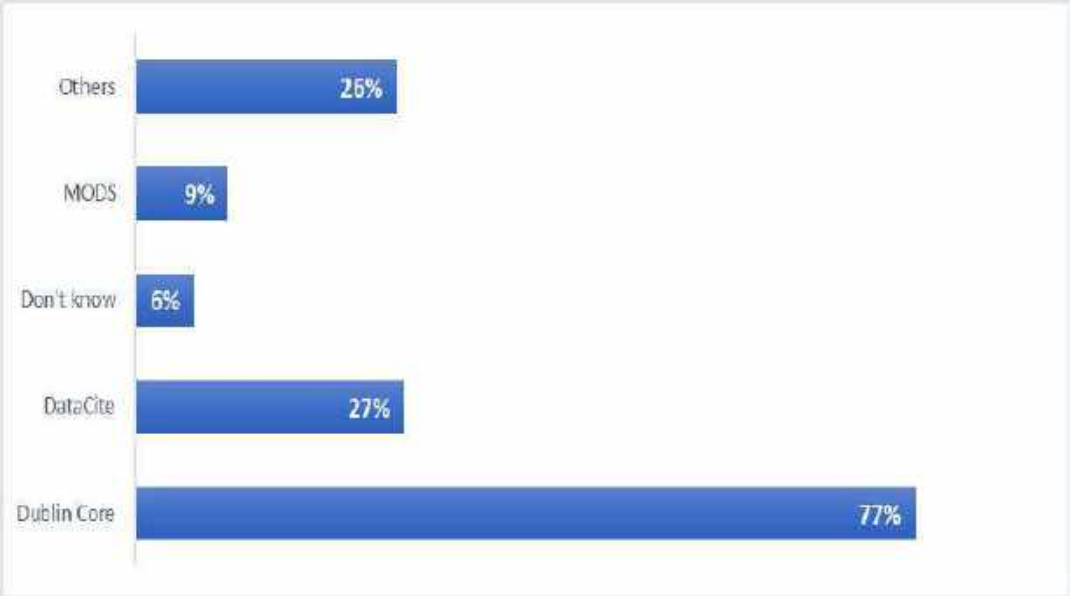

Figure 14: Metadata schemas available to use in the repository

OpenAIRE Guidelines

The OpenAIRE guidelines, which are more extensive and detailed than Dublin Core and include additional metadata elements such as funder and project IDs and access status, are becoming a widely used standard in Europe as they have been recommended by the European Commission (EC) as part of their open access policy. Many repositories in Europe (74%) have adopted the OpenAIRE Guidelines (Figure 15). It is worth noting that a significant number of repositories (167) are still using older versions of the Guidelines (which are less granular and don't include identifier schemes for authors, organisations or funders, and the COAR Controlled Vocabularies), meaning they do not meet the current EC requirements for metadata.

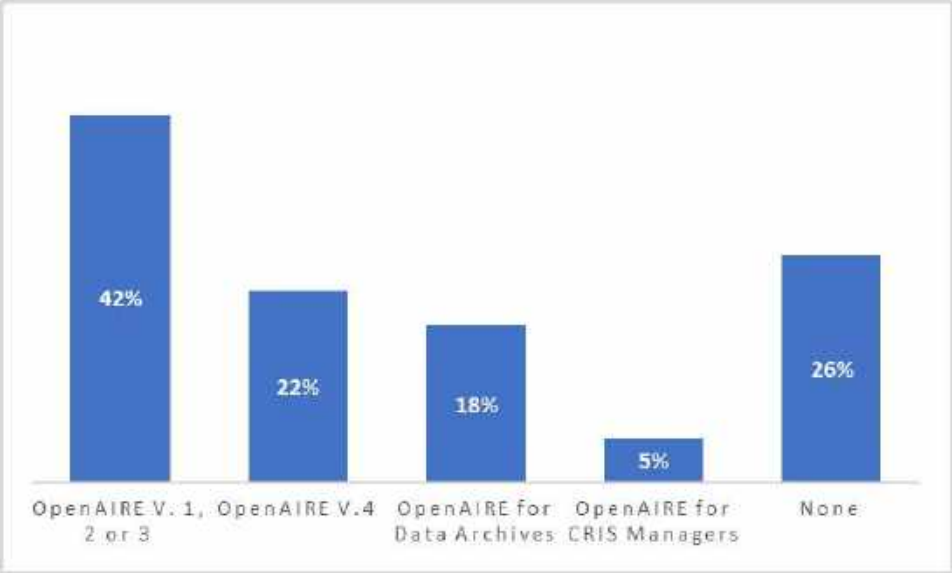

Figure 15: Support for OpenAIRE Guidelines

Licences

Almost all repositories (96%) offer users the option of choosing a specific licence, the most common of which are Creative Commons licences (91%). Some repositories offer several licensing options (Figure 16).

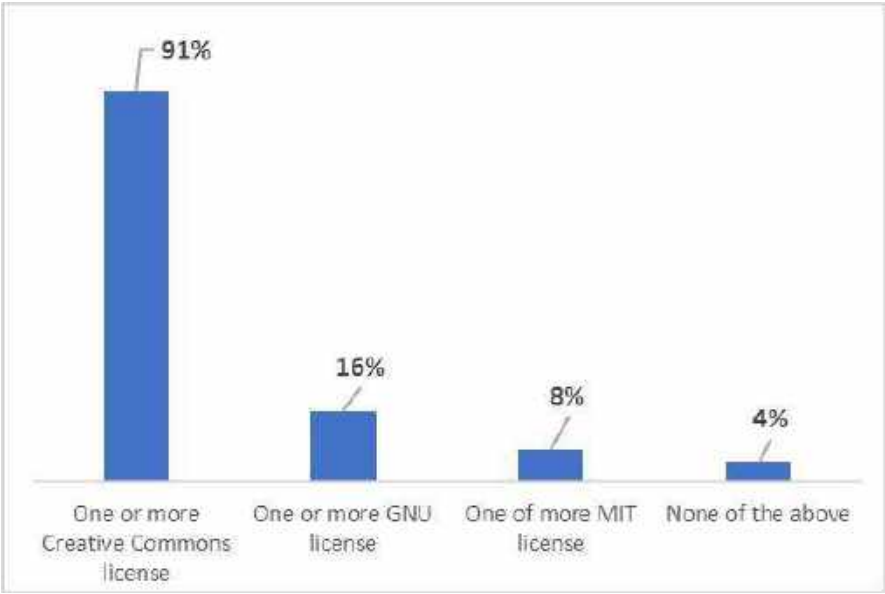

Figure 16: Licences available in the repository

Author IDs

ORCID IDs are quite widely supported, with 260 repositories providing a metadata field for ORCID in their records (66%), 71 support National IDs (18%), and other types of IDs are also supported by 78 repositories. 97 repositories do not support any type of author ID, which represents about 25% of respondents (Figure 17).

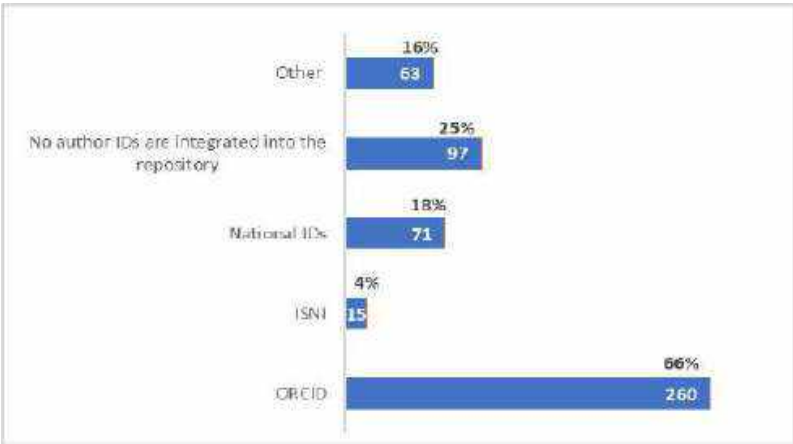

Figure 17: Authors IDs supported by the repository

Resource Persistent Identifiers

Many repositories assign at least one type of persistent identifier (PID) to the resources deposited, with the most common one being DOIs (Digital Object Identifiers) - 46%, followed by Handles (44%). 67 repositories support both Handles and DOIs. In the "other" category, most indicated that they are using an URN (Uniform Resource Name) or ISSN. About 10% of repositories do not assign / support any type of PID for the resources in their repository.

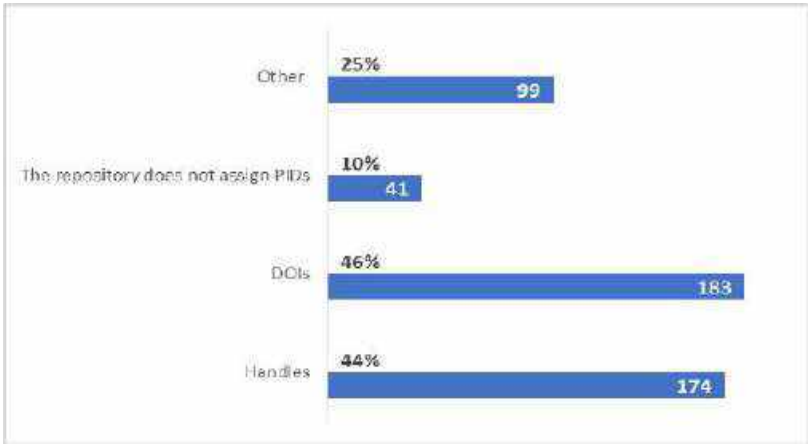

Figure 18: Persistent identifiers for resources assigned by the repository

Other services

Preservation

Approximately 63% of respondents (229) have a formal preservation policy in place at their repository, while 37% (136) indicated they have no preservation policy (Figure 19). In the comments, some respondents indicated that they were in the process of developing a policy (15); and several respondents noted that, while they don't have a formal policy, they do have a variety of preservation practices and procedures in place, including making back-up copies/mirroring content elsewhere. Some repositories are integrated with broader institutional preservation systems.

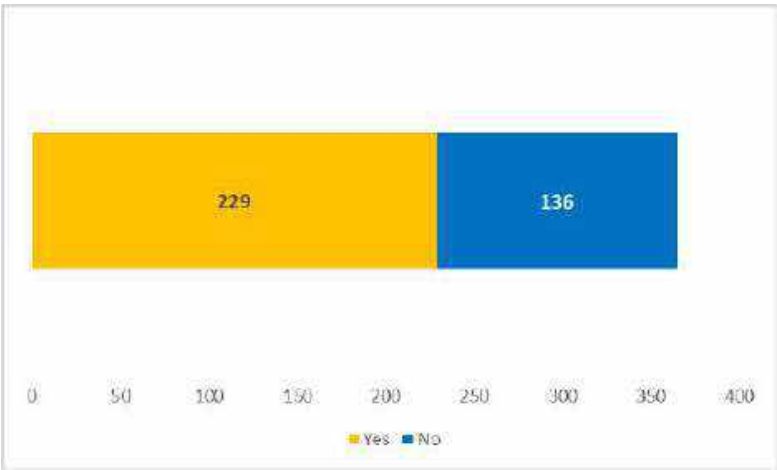

Figure 19: Repositories with a preservation policy

Usage statistics

Most respondent repositories (73%) are collecting usage statistics, with several using more than one usage statistics service. Only 33 repositories (about 10%) indicating they do not collect any type of usage stats. Most common is the use of the local repository statistics functionality, which is provided by the software platform.

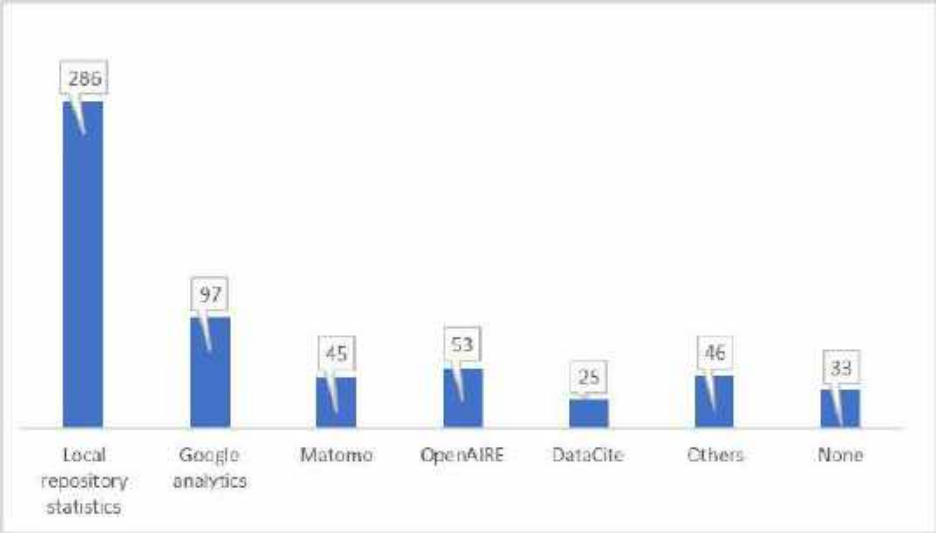

Figure 20: Type of repository statistics services used by the repository

Curation

Most repositories apply some level of curation upon deposit of a new resource. Metadata validation is the most common (checking that it is correct and/or complete), followed by mediated deposit (repository staff deposit on behalf of the researchers) and content validation (checking file formats and copyright) (Figure 21). In the "other" category, respondents listed things such as review for compliance with other deposit guidelines, checksum validation, and ethics review. Repositories do not undertake editorial review, but rather ensure resources are described and formatted properly.

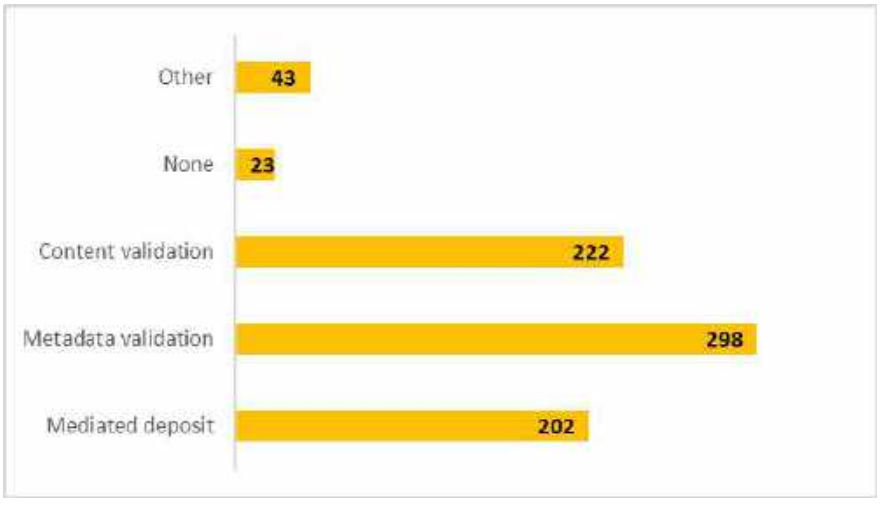

Figure 21: Curation process undertaken by the repository

Certification

23% of respondents said that the repository has undergone some type of certification (Figure 22), with CORE Trust Seal being the most common, followed by DINI and Data Seal of Approval. No significance difference in certification rates was found across repository types, with a slight increase for research data repositories. 19 respondents indicated compliance with national aggregator requirements or OpenAIRE as "certification" (which is not so much of a certification, but rather validation of the use of the OpenAIRE guidelines) (Table 3).

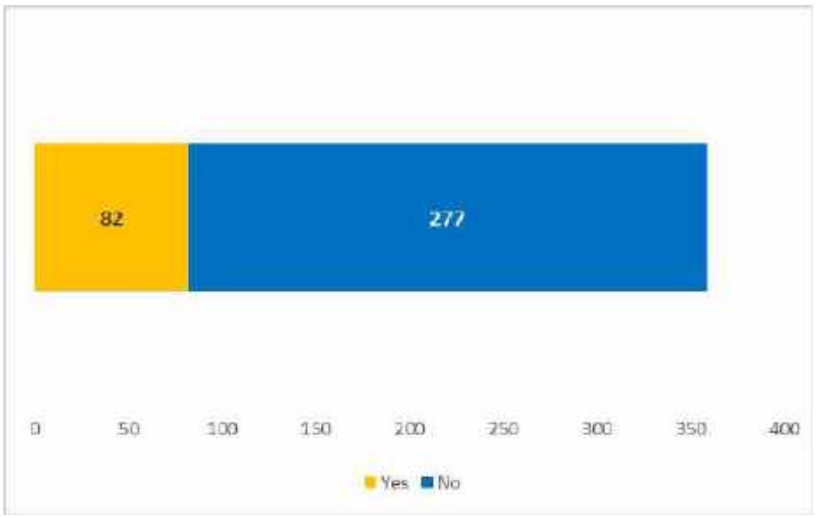

Figure 22: The repository has undergone some type of certification

Table 3: Type of certification undergone by repository

CoreTrustSeal	**20**
DINI	**14**
National Aggregator Compliance	**14**
Data Seal of Approval	**7**
OpenAIRE Compliance	**5**
ISO	**3**

Other value added services

Numerous other services beyond the ones mentioned above were described by respondents. Most common is the integration of repositories with other institutional services, such as a CRIS (current research information system), academic profile pages, or university websites.

A significant number of respondents also indicated that repository resources are reused by other types of external systems such as aggregators and discovery systems, but are also integrated into customised collections at the national level, reused for research assessment exercises (e.g. REF), and incorporated into national education curriculum.

Enhancement of repository records using metadata from other systems (e.g. using ORCID, Crossref records) is also common, as is the export of repository metadata to other systems. Other tools/functionalities such as using the CORE recommender system, digitization services, plagiarism detection, and request-a-copy were also mentioned.

Training was also widely referred to, especially by data repositories, which often provide training to researchers on how to format their data and how to complete data management plans. Some repositories offer assistance for authors to navigate copyright and other licensing issues.

Main funding sources

Institutional funding represents the predominant funding source for repositories, with 77% of respondents indicating their main funding source was their institution. 13% receive external project funding (Figure 23). Very few repositories (5, or just over 1%) charge a fee for depositors, and after further examination, these fees were only applied for certain types of deposits (i.e., unusually large data sets that require significant storage capacity). Most repositories rely on a single funding source, with only a few that receive funds from more than one source (institution and project funds mainly).

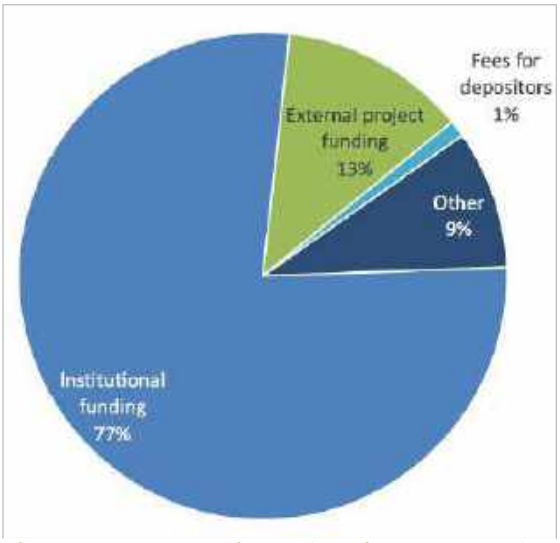

Figure 23: Predominant funding sources for the repository

Staffing

After removing several outlier responses with unrealistically high numbers (we presume these questions were misinterpreted by some), the average number of staff per repository was found to be just under 3 full time staff members (FTE). The staffing for repositories is spread across several positions: repository managers, technical support, metadata and content curation, and "other" positions. Close to half of the staffing of repositories (47%) is devoted to metadata and content curation, 27% to the repository manager position and 19% to technical support positions (Figure 24). Over half of respondent repositories have 2 or less full time employees (Figure 25).

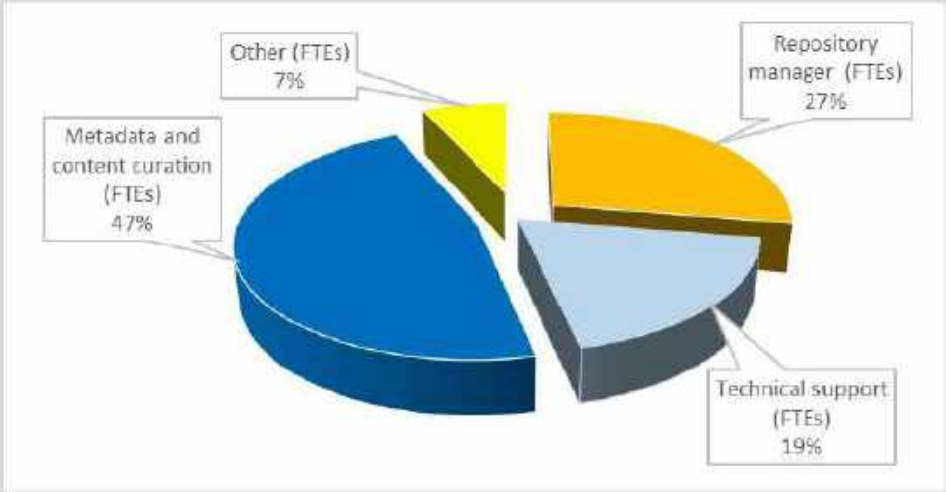

Figure 24: Distribution of staffing in repositories

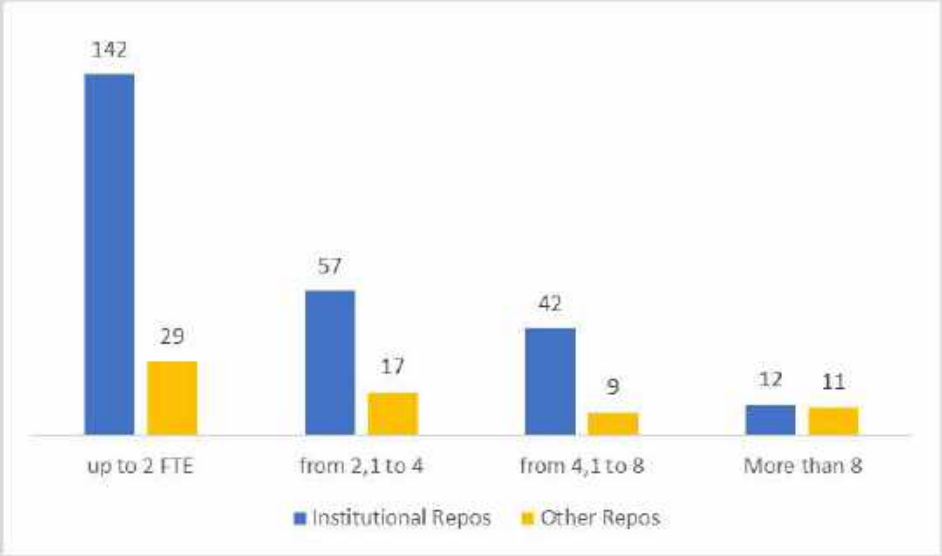

Figure 25: Number of staff members per repository

Sustainability

66% of respondents indicated that the repository was "very" sustainable and 31% indicated the repository was "somewhat" sustainable for the next three years. Some respondents noted that the repository has been a well-established service already for many years and is well used and well supported by their institution.

Several respondents provided more information about their sustainability challenges, grouped into several categories:

- Time and resource requirements to properly curate metadata and content
- Replacement of repositories with CRIS systems, which do not fully support the needs for managing a variety of content types
- Complexities of regular software upgrades
- High cost of employing outside companies to support software upgrades and ongoing maintenance of the system
- Lack of expected functionalities of the repository platforms
- Understaffing
- Project-oriented funding model

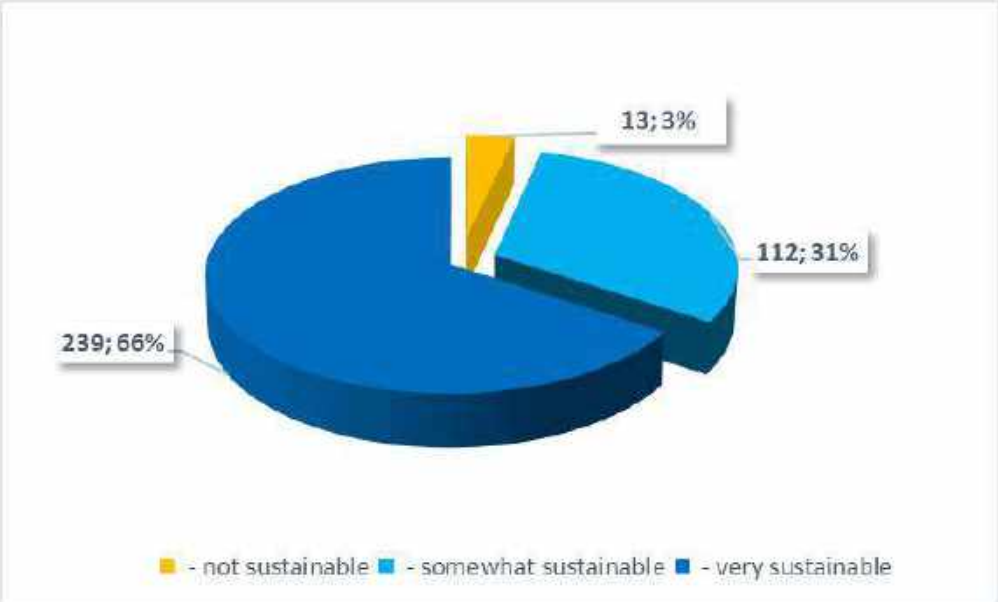

Figure 26: Respondents perceptions of repository sustainability

Together 97% of respondents (351) felt their repository was either "very" or "somewhat" sustainable, with only 13 respondents (3%) indicating that it was "not sustainable" (Figure 26). The 3% of respondents that felt their repository was unsustainable came from different countries and repository types, so no geographic generalisations could be inferred.

Challenges

Respondents were asked to rank a number of challenges to their repository operations. Software upgrades ranked as the biggest challenge (Figure 26), with 39% of respondents indicating this was a big challenge and another 38% indicating that it was somewhat of a challenge. This was followed by employing skilled staff, with 28% asserting that this was a big challenge. and then underfunding which was flagged by 26% of respondents. It should be noted that more than 50% of respondents indicated that all the five issues proposed in the survey were either "a big challenge" or "somewhat of a challenge".

We received numerous comments related to this survey question that fell into several other types of categories: lack of needed functionality of the software, policy trends moving away from repositories (e.g. a growing emphasis on gold open access), and complexities of the growing diversity and size of collections.

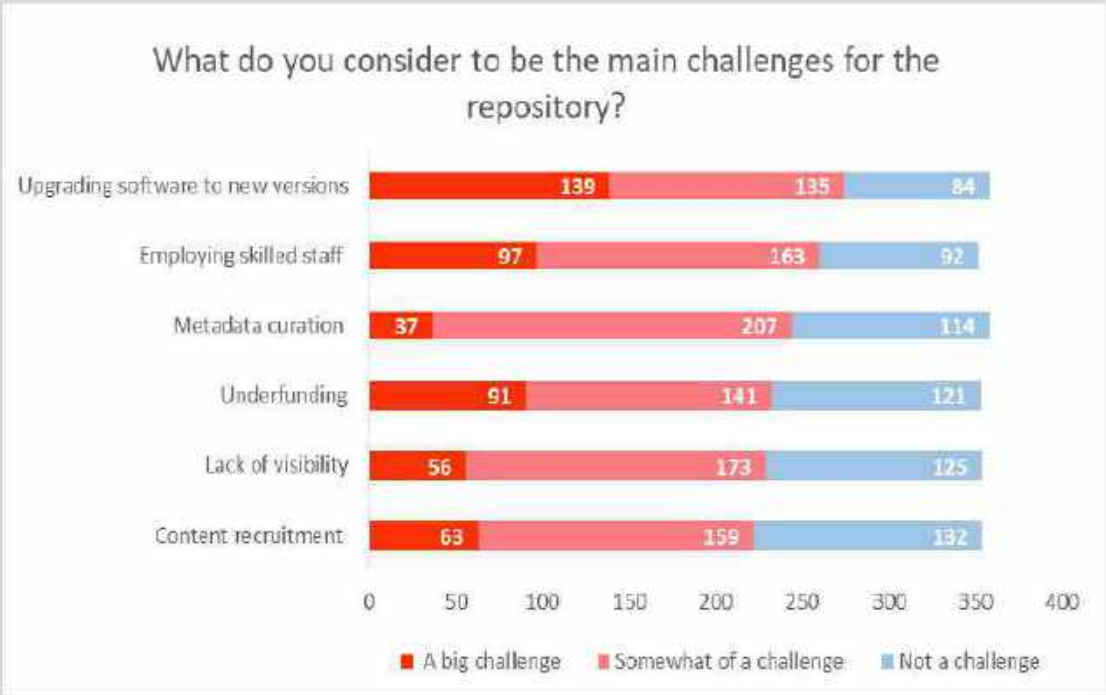

Figure 27: Challenges for repositories

Solutions / strategies

In terms of helping to address existing challenges, respondents ranked several proposed activities as "somewhat" or "very" helpful (Figure 28). The majority of respondents ranked all options provided as very helpful or somewhat helpful. Advocacy for repositories was ranked highest, with 58% indicating it would be very helpful and 34% saying it would be somewhat helpful. Community of practice for technical support was also seen as very helpful with 92% of respondents indicating this would be very or somewhat helpful. Greater national and regional coordination for repositories and training for managers was also considered as very and somewhat helpful for the vast majority of respondents (86% and 85% respectively). A national or regional platform for hosting was ranked lowest, but still was considered helpful by just over 50% of respondents (this could be explained by the fact that several countries already have a national platform so it is not needed in those jurisdictions).

Some respondents provided other suggested solutions including the development of tools and services to assist with ingest and curation, increased funding, and improving incentives for researchers to deposit.

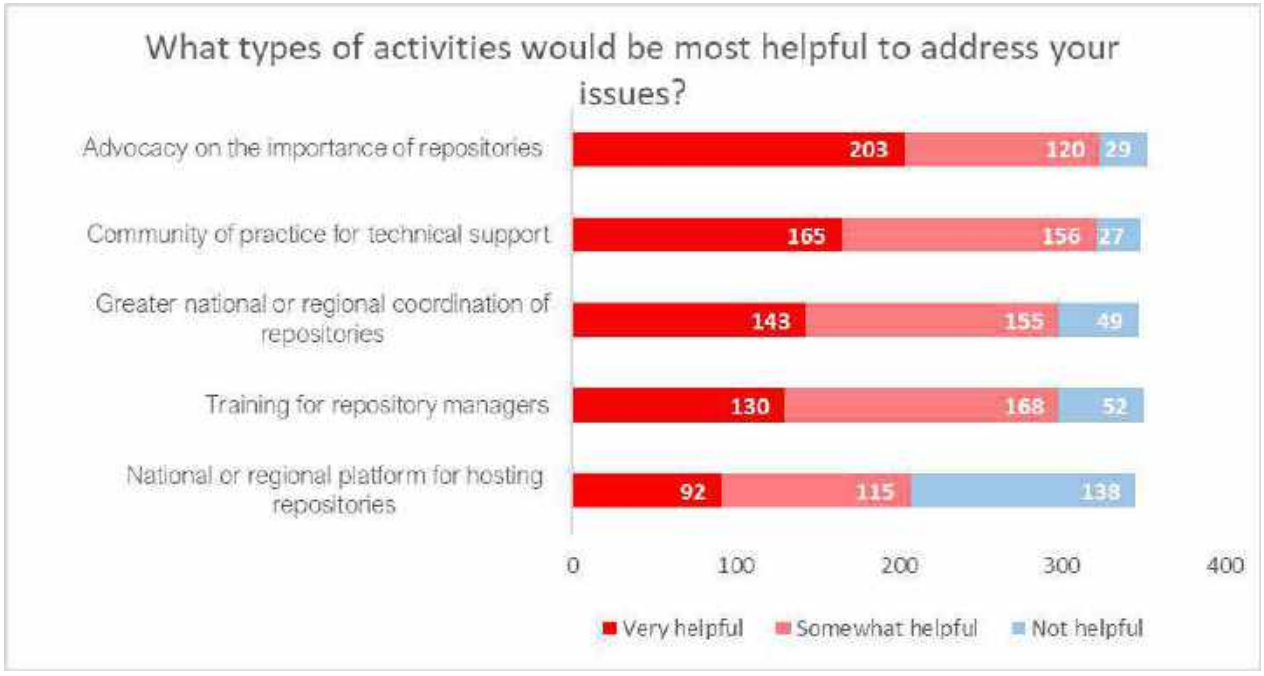

Figure 28: Activities that will help to address challenges

Analysis

The survey had 394 responses from 34 countries in Europe. While this is only a portion of the repositories in the region - OpenDOAR and Re3Data together list over 3,000 European repositories - we believe that the collected sample is somewhat representative of the European repository landscape, but perhaps slightly skewed towards publication repositories.

A strong and well-functioning network of repositories that provides human and machine access to the wide range of valuable research outputs is needed for Europe to reap the full benefits of open science. This will require repositories to be sustainable; user friendly and technically agile; and able to interoperate with a range of other value-added services. Below is a summary of our key findings and highlight the strengths and weaknesses in the current landscape.

Coverage

The majority of the repositories that responded to the survey were institutional repositories. This may be partially due to the fact that the survey was disseminated by library-based organisations, but also reflects the reality that the vast majority of repositories in Europe are managed by universities/university libraries or research centres. Institutional repositories are generally quite sustainable, as they are hosted by long-lived institutions, who have committed budgets towards this activity. While OA repositories are open for anyone to access their content, most repositories focus on collecting research outputs created by members of a specific community, which typically fall into one of a few categories: institutional, national, international, and domain repositories. With the current prevalence and variety of repositories in Europe, all researchers will have at least one repository in which they can share their research outputs.

Collections

European repositories collect a variety of content types ranging from journal articles (author accepted manuscripts and publisher versions), e-theses and dissertations, research data, as well as a range of other materials. Institutional, international, and national repositories tend to collect a diversity of content types and disciplines; domain repositories usually focus on a specific content type in addition to their disciplinary coverage. Collection sizes vary significantly across

respondent repositories, with an average size of 64,859 items for institutional repositories and 386,088 items for the other types of repositories (data, domain, generalist and national repositories) and collection sizes range from less than a thousand items to millions of resources. The largest respondent repository was Europe PMC, which contains over 8 million full text records. Even if we consider a low estimate of around 1,500 active repositories in Europe with an average of 65,000 items per repository; this would mean European repositories collectively provide open access to close to 100 million items. Collectively, this represents a significant amount of content if we consider that the large knowledge graphs that aggregate from many different content providers contain between 200-300 million objects[3],[4].

Repositories support bibliodiversity in the ecosystem. They do not charge for access or for research to deposit and they collect and preserve a range of content types in many domains and disciplines. The most predominant content types in repositories are journal articles, theses and dissertations, and conference proceedings. Research data is also common, followed by a long tail of other types of scholarly and educational materials. Some repositories contain preprints, but they still represent a small portion of items in the European network. Repositories, therefore, are well placed to support the expansion of open science practices across Europe and the reformation of research assessment, which places a greater emphasis on inclusiveness, diversity, and transparency.

Multilingualism

There is a growing recognition in Europe and beyond of the need to support and encourage publishing in local languages, as this ensures that the public has access to the research (which they often fund). The survey found that repositories play an important role in preserving and disseminating content in a variety of languages, especially local languages. 17% of respondent repositories collect content in only one language, almost exclusively represented by repositories in predominantly English-speaking countries (UK and Ireland), meaning most repositories collect content in at least two languages.

The survey also found some diversity in the languages of resources across the European repository network, with at least 29 languages represented in total. That said, the predominant language for over 50% of respondent repositories was

[3] On Nov 24, 2023, OpenAlex had 246 million objects in its aggregation: https://help.openalex.org/
[4] On Nov 24, 2023, OpenAIRE has 239 million objects in its aggregation: https://graph.openaire.eu/

English, even for many repositories from non-English speaking countries. There are 24 official languages in the EU, however, over 200 languages are spoken across the continent[5]. So there are still many languages that do not seem to be well represented in the European repository network. Repositories tend to collect resources in just two or three languages, with either the main local language being most predominant, or second most predominant after English. As English is the lingua-franca for research, especially in the STM (science, technology and medical fields), these results are not unexpected.

As there are fewer international venues for disseminating non-English content, institutional repositories are playing a role in this respect for some of their local communities. In a few cases, repositories publish metadata and abstracts in two languages (usually the original language of the resource and English), which can lead to better discovery in (the predominantly English-focused) indexing and discovery services. We know from anecdotal information[6] that repository platforms have typically been developed with English in mind and do not support all languages equally. Therefore, managing non-English content can involve extra efforts for those repositories such as translating the interface of the platform and metadata curation to correctly assign language codes, especially for languages that use non-roman characters. This may also partially explain the predominance of English in repositories.

Services

Along with their primary role of collecting and providing access to research outputs, repositories are active participants in a broader scholarly ecosystem, feeding their metadata and resources into various types of networks and services that repurpose the content and/or combine it with others. Almost all repositories offer certain baseline services: metadata checking, deposit support, back-up copies, and usage statistics. The vast majority of repositories expose their metadata (and sometimes full text resources) to external discovery services using OAI-PMH protocol. In addition, repository records are increasingly indexed / visible in other external systems, such as DataCite (because they are minting DOIs).

The next generation repository envisions repositories as more than institutional services but as the foundation for other services built on the collective contents

[5] https://www.tomedes.com/translator-hub/european-languages

[6] See the discuss and analysis by the COAR Task Force on Supporting Multilingualism and non-English Content in Repositories

of repositories[7]. The impact and value of repositories, therefore, can be demonstrated by both local usage and download statistics, as well as the downstream reuse of repository resources in other contexts. In this respect, repository collections are increasingly being repurposed and reused in innovative ways. In particular, the integration of repository records into institutional or national systems, such as CRIS systems (especially common in the UK), academic profile pages, university websites, and other internal research administrative tools is widespread as is the reuse and repurposing of repository content into other collections, such domain collections, specialised portals, and education curricula is also becoming quite common.

Metadata and persistent identifiers

The use of PIDs and comprehensive and standardised metadata is a fundamental requirement for the discovery and reuse of repository resources. The vast majority of repositories support Dublin Core metadata, which has typically been the default for repositories. This, therefore, is the baseline of interoperability for repositories in Europe. In addition, there has also been quite widespread adoption of the OpenAIRE guidelines, which are more granular and require additional metadata elements to be added to repository records.

In terms of persistent identifiers (PIDs), most repositories are now using PIDs for their resources (either handles or DOIs). This is a positive development because it ensures a certain level of permanence for the resources in repositories (for example, if URLs change when a repository changes platforms or upgrades to a new version). Other types of PIDs are also increasingly supported by repositories including author IDs, funder, and institution IDs, and so on, bring additional benefits: enabling the analysis and tracking of research outputs according to the funder, university, or author; and providing an opportunity for repositories to integrate metadata from those external systems to enhance their local metadata.

Despite the fact that a repository supports certain metadata schemas and PIDs does not always equate to the collections having high quality metadata. While most repositories do support standardised and granular metadata schemas, they often rely on the author to fill in the metadata fields. Since authors may not be

[7] https://www.coar-repositories.org/news-updates/what-we-do/next-generation-repositories

aware of the standards, this often leads to lower quality metadata records[8]. While most repositories do undertake basic metadata curation and checking, this may not be sufficient to optimise discovery and reuse of repository resources. There are opportunities to improve the quality of metadata - either through data curation activities at the repository, or by introducing machine extraction of metadata information - but this may require greater commitment in terms of staff and technical resources at the repository.

Technologies and functionalities

Respondents seemed overall satisfied with the repository platforms they are using and their functionalities, with the exception of a few respondents who felt that the platform was not able to respond and adopt new technologies quickly enough. Some respondents mentioned their current repository was not fit for purpose (in particular several who are using a CRIS system as their main repository). However, the requirement to continually upgrade repository software to newer versions is a challenge. Upgrading repository software is not a trivial task and requires significant technical resources, often taking several months to complete. Adding to this, over 50% of respondents indicated that there have been changes made to the standard code base of the platform at their repository, contributing to the complexity of upgrades and making it more difficult to transition to newer versions. More than 60% of respondents either upgraded in 2022 or plan to upgrade in 2023. 20% of respondents will undergo an upgrade two years in a row: 2022 and 2023. Notably, DSpace, the most widely used platform (representing just under half of the repositories in the survey), has announced that their support for earlier versions will be ending in 2023, and this may account for a higher number of respondents who are upgrading now. Open-source platforms regularly develop new versions to remain competitive in the market and support technological expectations of users. Yet, this introduces technical demands that consume a large amount of staff time and may be diverting resources away from other important repository operations such as engaging with researchers on campus, improving metadata quality, or engaging with value added services.

[8] Numerous studies have found that mediated deposit by librarians and repository staff improve the quality of the metadata in the repository record. See for example: Roy, Bijan Kumar, "Institutional Digital Repositories: a systematic review of literature" (2021). Library Philosophy and Practice (e-journal). 4855. https://digitalcommons.unl.edu/libphilprac/4855

Certification

Most repositories do not make use of existing certification frameworks, either because the assessment process is too resource intensive or existing requirements are deemed as unattainable. An alternative, perhaps lighter-weight self-assessment framework may be more widely applicable for the majority of repositories, perhaps based on the "COAR Community Framework for Good Practices in Repositories"[9] (which has already been adapted into a self-assessment tool in the Japanese context). Certification patterns are aligned at the national level - that is, there are a few countries where certification rates are much higher - therefore certification may be most effectively propagated via national agencies or communities.

Sustainability and funding

Sustainability of repository operations was considered quite high by respondents, with only 3% of respondents indicating that there is a risk to their operations. This may be related to the fact that most repositories are affiliated with an institution - usually the university library - and therefore have a dedicated budget and stable staffing (as opposed to being funded by project grants or other short-term means). That said, 31% indicated that the repository was only "somewhat" sustainable for a number of reasons related to several challenges, most notably the problem of managing software (both internally at the institution, or the high costs of paying an external provider).

Just over 50% of repositories have staffing levels of under 2 full time employees, after combining the time from all the different staff positions. This could be considered quite low, given the operational requirements of a repository (although this depends on a number of other factors, such as external hosting and size of repository). Increased staffing at these repositories could help to address many of the challenges being experienced and ensure there is widespread adoption of good practices and next generation repository functionalities. Shared infrastructure models, which have already been adopted in a few countries, are another approach that offer economies of scale and could relieve some of the burden from individual institutions.

[9] https://www.coar-repositories.org/coar-community-framework-for-good-practices-in-repositories/

Conclusions

Collectively, European repositories acquire, preserve and provide open access to tens or possibly hundreds of millions of valuable research outputs and represent critical, not-for-profit infrastructure in the European open science landscape. They are used for sharing articles that may be paywalled in published journals, but also for providing access to a variety of other types of research outputs including theses/dissertations, conference papers, research data, preprints, code, and so on. They will be critical infrastructure as Europe collectively advances open science and research reform that incentivises the sharing of all valuable research outputs.

Many repositories are based at universities making them quite sustainable and, by every indication, their collections are being well-used by the research community and beyond. Given the general concerns about fragility of open science infrastructures, a distributed approach, with national and regional nodes, seems to be a viable model for other types of scholarly communications infrastructures.

> *"We see massive use of our thesis repository – no way people would get to read these theses were we not making them available in this way.*
>
> **- Survey respondent**

The number and range of value added services to which repositories are contributing demonstrates that European repositories have been progressing towards the vision of the next generation repository, which is about moving beyond the repository as an institutional service, to the networked repository that is an integral part of the broader ecosystem. However, to fully achieve our collective vision, there is still work to be done. The survey has exposed a number of important areas where the current repository landscape could be strengthened. In particular, we found that repositories struggle with three main challenges: (1) maintaining up-to-date, highly functioning software platforms, (2) applying consistent and comprehensive good practices in terms of metadata, preservation, and usage statistics; and (3) gaining appropriate visibility in the scholarly ecosystem. These challenges can be traced to several interrelated underlying factors:

Managing local software: Open-source software is the obvious preference for most repositories, as it enables the institution to participate in the governance of the software, make changes to the code to support local needs (e.g., language, functionalities, integrations with other systems), and belong to a community of practice with other software adopters. Yet, managing software locally requires local technical expertise and a significant time commitment. Many repositories have difficulty keeping up with the newest version of their software platform, which can have an effect on the service provision, as requirements change and user expectations evolve. In addition, there is an inherent tension in the repository ecosystem where - on the one hand - there is a need to ensure widespread interoperability and maintain ease of upgrades by not introducing special functionalities - and on the other hand - being responsive to the needs of various local and national communities that request certain tailored services (for example, local languages). Maintaining this balance can present a challenge for repositories, as they seek to provide a high quality service to their local communities while maintaining a modern repository platform.

Staffing levels: In terms of staffing, repositories have quite low numbers. This can contribute to the problems identified above with not having the capacity to upgrade to new versions when needed, but also can result in only a basic level of support for other services, such as user support, metadata curation, and awareness of the repository in the community. As the needs of the user community expand and evolve with open science becoming mainstream, there will be an increasing strain on repository staff. Low staffing levels are due to the fact that repositories have not been a high priority service for universities and that they are also competing with the commercial sector for skilled technical staff.

Distributed nature of repositories: One of the great strengths of the repository ecosystem in Europe and beyond is its distributed nature. This contributes to the sustainability of the network as it is collectively funded by many universities and research centres, as well as circumvents a situation where there are only a few points of failure. However, this highly distributed environment also creates a situation where repositories can have low visibility, are isolated, and are working in silos with little opportunity to share expertise and learn from colleagues. To some extent, repositories are replicating services across many institutions. That said, over the several years, a number of countries have adopted a more coordinated approach to repositories through national discovery services, shared infrastructure models, and hosting communities of practice. This allows institutions to benefit from some economies of scale and address some of the challenges of the distributed environment.

Opportunities and Next Steps

Despite the challenges, the current climate offers exciting opportunities for repositories. Many funders are actively promoting the repository route for articles because of their role in supporting equitable access to content (i.e., no fees to access or deposit). The value proposition for open science is growing and repositories are increasingly recognised as the main mechanism for collecting and providing access to a wide range of research outputs. Add to this, the nascent, but growing, interest in the publish-review-curate model in which repositories have a central function[10], and it seems they are well placed to expand their current role in the ecosystem.

To support this evolving role for repositories, OpenAIRE, LIBER, SPARC Europe and COAR have identified three areas in which we can work together to help advance and strengthen repositories in Europe:

1. Highlighting the value proposition for repositories and advocating for the critical role of repositories in Europe
2. Propagating best practices for repositories across the continent
3. Assisting with the creation and coordination of national networks

In the coming months, our organisations will develop more concrete plans for advancing each of these areas.

Data Availability Statement

The anonymised data that support these findings and the survey questionnaire are openly available in Zenodo.

DOI: https://zenodo.org/doi/10.5281/zenodo.10213483.

[10] From cOAlition S: To illustrate how a scholar-led communication system can (and already does) work in practice and supports the principles of Open Science, we highlight the Publish, Review, Curate (PRC) model, which we find particularly promising. https://www.coalition-s.org/wp-content/uploads/2023/10/Towards_Responsible_Publishing_web.pdf

Current State and Future Directions for Open Repositories in Europe

Results of Survey of Open Repositories in Europe

December 2023

OpenAIRE LIBER SPARC Europe COAR