European Research Council
Established by the European Commission

European Commission

Study on the readiness of research data and literature repositories to facilitate compliance with the Open Science Horizon Europe MGA requirements

Final Report

Expert Group for this study:
Najko Jahn, https://orcid.org/0000-0001-5105-1463
Mikael Laakso, https://orcid.org/0000-0003-3951-7990
Emma Lazzeri, https://orcid.org/0000-0003-0506-046X
Peter McQuilton, https://orcid.org/0000-0003-2687-1982

Study on the readiness of research data and literature repositories to facilitate compliance with the Open Science Horizon Europe MGA requirements

ERCEA - The European Research Council Executive Agency

Contact: Anna Pelagotti (Anna.PELAGOTTI@ec.europa.eu)
Dagmar Meyer (Dagmar.MEYER@ec.europa.eu)

Manuscript completed in November 2022.

DOI: 10.5281/zenodo.7728016

LEGAL NOTICE

This document has been prepared for the European Research Council Executive Agency (ERCEA). However, the views set out in this document are those of the authors only and do not necessarily reflect the official opinion of the ERCEA. The ERCEA does not guarantee the accuracy of the data included in this document. Neither the European Commission nor the ERCEA nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

Contents

Summary

This report and associated repository inventory represent the output of a study conducted between March and October 2022 by a group of independent experts and commissioned by the European Research Council Executive Agency (ERCEA). In this study we assess and analyse the readiness of research data and literature repositories to facilitate compliance with the Open Science requirements in the Horizon Europe Model Grant Agreement (HE MGA) (European Commission, 2022a). This study also takes current repository choice practices of ERC-funded researchers into account.

This study aims to:

- create a better general understanding of the availability and nature of repositories in different fields of research, for both literature and research data;
- identify trusted repositories across different fields of research, and highlight those that are most widely used by ERC-funded researchers;
- assess to what degree the identified trusted repositories facilitate compliance with the HE MGA requirements related to Open Science, in particular with regard to the metadata of deposited research outputs;
- enable the ERC Scientific Council to provide well-founded guidance to ERC grantees as to which repositories will allow them to fulfil the Open Science related obligations of their HE grant agreement.

In this study we analysed 220 repositories, and, via a structured methodology, we identified 165 trusted repositories and tested their readiness to facilitate the compliance with the HE MGA Open Science requirements.

We show that it is not straightforward to assess whether a given repository is suitable to facilitate compliance with the HE MGA requirements. This is mainly due to varying interpretations of definitions and requirements, whether information on repository specifications is publicly available, and the high level of technical expertise needed to assess all requirements.

We highlight that repository registries, such as FAIRsharing, re3data or the CoreTrustSeal (CTS) website, are not sufficient on their own to assess the readiness of repositories to facilitate compliance with the HE MGA requirements, as the definition of what constitutes a trusted repository is subtle and varied and needs to be carefully interpreted and applied to repositories. This is also the case for related concepts such as community endorsement or for policy requirements in terms of preservation, curation and security of the repository contents.

We found that current certification is not always in line with funder requirements as many repositories that hold a certification do not meet all the essential characteristics criteria for trusted repositories set out in the HE Annotated Model Grant Agreement.

Only three repositories identified as trusted in this study also fulfil all the mandatory requirements for metadata, whereas none meet all the mandatory and recommended metadata requirements as they are set out in the HE MGA.

Two of such trusted repositories have international coverage. One of them is for books and the other one is of a ‘catch all’ type.

Repositories are defined as catch-all when they can host digital content independently of their nature (data, literature, etc.).

This study has given us the opportunity to deeply analyse a set of repositories, their technical characteristics, policy framework and usage information. From our work, we can see that data repositories, as opposed to literature repositories, appear to be more ready to facilitate the compliance of the grantees with the HE AGA definition.

This study and its related inventory aim to provide

- A general overview of the repository landscape in different research areas
- A tool for use by ERC grantees to easily find a repository that facilitates compliance with the HE MGA Open Science obligations, the inventory of identified trusted repositories
- A methodology for assessing repository readiness for facilitating compliance with the HE MGA requirements
- An analysis and related discussion of a set of selected repositories and their capacity to support ERC grantees to satisfy the HE MGA requirements in terms of Open Science.

This study and related inventory are not intended to provide

- A definitive and exhaustive list of repositories that facilitate compliance with the HE MGA Open Science requirements. Data collection for the initial inventory has been skewed towards the resources available during the study, and we expect the inventory to grow over time and to evolve, so as to be of even more value to ERC grantees.
- A tool to assess repository FAIRness. Although some FAIR-related aspects are part of the repository features that are assessed, a holistic FAIR perspective is not within the scope of this study.
- A formal certification mechanism

1. Introduction

Open Science relies on well structured, licensed, and contextualised literature and data that are being managed appropriately and deposited in infrastructures that serve this scope: repositories.

> A repository is an online archive, where researchers can deposit digital research outputs and provide (open) access to them. Repositories help manage and provide access to scientific outputs, such as publications, data, software, among others. They also contribute to the long-term preservation of digital assets. Repositories can be institutional, operating with the purpose to collect, disseminate and preserve digital research outputs of individual research organisations (institutional repositories, e.g. the repository of University X). They can be domain-specific, operating to support specific research communities and supported/endorsed by them (e.g. Europe PMC for Life Sciences including biomedicine and health or arXiv for physics, mathematics, computer science, quantitative biology, quantitative finance and statistics; Phonogrammarchiv for audiovisual recordings, the CLARIN-DK-UCPH Repository for digital language data or the European Nucleotide Archive or databases of astronomical observations operated by the European Southern Observatory, among others). There are also general-purpose repositories, such as for example Zenodo, developed by CERN.
>
> *Horizon Europe Annotated Model Grant Agreement, Page 155*
> *(European Commission, 2021)*

The number of literature and data repositories has significantly increased over the last two decades and as of November 2022 there are just under 6000 repositories listed in the Directory of Open Access Repositories (OpenDOAR[1]), with over 1000 manually curated national and international domain-specific and general-purpose (or general) repositories in FAIRsharing[2] and about 3000 data repositories registered in re3data[3]. Open repositories are very diverse when it comes to the type of digital objects they host, their curation, the kind of organisational context they operate in, and the kind of software and hardware that sustain them. While diversity has been key to enabling flexibility in repository implementation in different settings and for varying purposes, it can be challenging for researchers to understand and compare the different options available to them. This is further compounded by the lack of widespread adoption of formal certifications (e.g., only ~160 repositories had a valid CoreTrustSeal (CTS) certification as of April 2022[4]). There are no widely adopted categorisation frameworks that can be used to assess data and literature repositories, the metadata they provide, their Open Science or FAIR credentials, or licensing surrounding their data/literature and associated

[1] https://v2.sherpa.ac.uk/opendoar/
[2] https://fairsharing.org/
[3] https://www.re3data.org/
[4] https://www.coretrustseal.org/

metadata. Attempts such as Bioschemas[5] and the EOSC Marketplace[6] are beginning to address this but are yet to gain sufficient traction in the community.

Repositories may be specific to a discipline or digital object or can be domain/object-agnostic. Many focus on datasets, but there are a number that cater for other types of digital objects, such as publications, preprints, software, and algorithms. These repositories play a key role in ensuring greater transparency and preservation of research findings, as expected by funders, governments and learned societies. To support the findability, accessibility, interoperability and reusability of data, many repositories are beginning to incorporate the FAIR principles (Wilkinson, 2016) into their policies and implement the necessary technical enhancements.

Alongside this push to enhance repositories to meet the latest standards and best practices, the urge to assess their readiness to meet the need of the stakeholders has become key, especially since the use of repositories was linked to the possibility to meet the funders' obligations in terms of Open Science and FAIR Research Data Management. Different tools have been developed to assess the FAIRness of both data (FAIR Data Self Assessment Tool[7], SATIFYD[8]) and repositories (Devaraju et al., 2020), and recently the responsiveness and readiness of research data repositories in terms of the implementation of FAIR principles, including certifications, was the object of a study commissioned by the European Commission, Directorate General for Research and Innovation (European Commission, 2022b).

A central concept that intersects both social and technical aspects of repository management is trust, giving users and their funders confidence in responsible long-term management and sustainability of hosted content. Web services and their content run the risk of being ephemeral unless precautions are in place to ensure the secure preservation of the content and commitment to maintain the technical architecture to modern standards. Several organisations have proposed criteria for the assessment of the quality of repositories and their recommendation, but there is no consensus as to what constitutes a 'good' repository. Examples are the domain-agnostic CoreTrustSeal (CTS), the nascent TRUST principles (Lin, Crabtree & Dillo et al., 2020), the domain-specific ELIXIR Core Data Resources (Drysdale, McEntyre, Durinx and Blomberg, 2018), journal publisher-specific guidance (Cannon et al., 2021) and funder-specific guidance such as that from the USA National Institutes of Health (NIH) (Office of The Director, National Institutes of Health, 2020). These different efforts attempt to define criteria and share many characteristics. However, while worthy, they have failed to lead to widespread adoption of a common set of principles or metadata standards. There are many reasons why this is the case: economic barriers (repositories must pay to have CTS certification), time and cultural factors (the 'nascent' TRUST principles are gaining adoption but are yet to spread across the many disciplines across data and literature repositories), or diversity of the research contexts (while domain-specific efforts may do well in their domains they can understandably fail to crossover to other disciplines, due to the different best practices, standards and research outputs peculiarities).

[5] https://www.bioschemas.org
[6] https://marketplace.eosc-portal.eu/
[7] https://ardc.edu.au/resource/fair-data-self-assessment-tool/
[8] https://satifyd.dans.knaw.nl/

This creates a challenging situation for research funders. It is not enough to simply instruct grantees to deposit publications, data and other digital objects into any repository, yet listing several specific criteria for repository functionality makes the process overly complex and time-consuming for researchers. Horizon Europe grantees are mandated to use trusted repositories in order to comply with Open Science obligations. However, there is no quick way to identify whether such repositories exist. This work aims to alleviate this situation by taking the definition of trusted repositories from the Horizon Europe Annotated Model Grant Agreement (HE AGA) (European Commission, 2021) alongside the requirements for specific metadata that concern the Open Science and FAIR data management that Horizon Europe beneficiaries have to comply with and apply them to repositories for literature and data. The aim of this work is to enable researchers to make an informed, evidence-based decision as to where to deposit their outputs. This study also aims to provide a list of repositories that would allow researchers to comply with the *majority* of the HE MGA requirements, and highlights research areas where no such domain/discipline-specific repositories exist (or where there are no repositories suitable for ERC-funded researchers).

Based on this context, we set out to deliver the following:

General overview of the repository landscape

- Presentation and discussion of the main characteristics of available repositories, highlighting differences between repositories across different domains. This overview distinguishes between repositories for publications and for research data.

Inventory of identified trusted repositories for different fields of research

- Inventory covering all three ERC domains (Life Sciences, Physical and Engineering Sciences, Social Sciences and Humanities), subdivided by (suitably chosen groups of) ERC evaluation panels[9]. The inventory also indicates when a repository is open to content from any scientific area. Although the focus is on domain-specific repositories, other repositories, such as institutional repositories are not excluded *a priori*.
- Presentation and motivation for the criteria used for inclusion in the inventory, alongside the methodology employed for the identification of repositories that fulfil those criteria
- Discussion of the most frequent / most important reasons for a repository to not fulfil the criteria to be considered a trusted repository

Assessment of repository readiness to facilitate compliance with the Horizon Europe MGA requirements

- Indication of the extent to which the trusted repositories included in the inventory facilitate compliance with the HE MGA requirements related to Open Science, in particular with regard to the metadata of the deposited research outputs
- Indication for each individual requirement in the HE MGA whether the repository allows ERC-funded authors to comply
- Identification of research areas where no discipline-specific or domain-specific repositories with the required qualities are available

[9] https://erc.europa.eu/news/new-erc-panel-structure-2021-and-2022

- Extension of the assessment to repositories that are heavily used by ERC-funded researchers but narrowly miss the criteria for being included in the inventory
- Indication of:
  - options for metadata access (e.g., through an API)
  - metadata standard used
  - compliance with the OpenAIRE guidelines[10].

This report is structured as follows: In the next section we present a general overview of the landscape for both literature and data repositories, including, for the latter, specific domain peculiarities. The next three sections have the following contents: discussion of the basic concepts behind the definition of trusted repositories included in the HE AGA, and of the HE MGA metadata requirements; presentation and discussion of the methodology used to select repositories to be included in the study and of those to form the repository inventory; analysis of the aggregated data derived from the study. In the final section we provide our conclusions from the work that has been done.

This study also includes the following annexes that represent the underlying data of the study and useful list and documentation for the ERC grantees:

- ANNEX 1: The inventory of identified trusted repositories as a spreadsheet, with separated sheets for data and literature repositories
- ANNEX 2: The survey questions used to collect information about the repositories included in the study that led to the inventory realisation
- ANNEX 3: The curated results of the survey and information collection that was used as source for the study and related analysis

[10] https://guidelines.openaire.eu/en/latest/

2. General overview of the repository landscape

Since the turn of the century, there has been an explosive growth in data production across the sciences. It can be hard to grasp the scale of this growth. As an example, 90% of all the data in the world today was created in only the last 2 years[11,12]. This poses a considerable data management problem. It has been estimated that approximately 80% of the data produced are lost in 20 years' time (Vines, Albert, Andrew et al., 2014). How can we preserve data in a more sustainable manner? How can we ensure the FAIRness of this data? Open, FAIR repositories, at an institutional, national or global level provide a solid foundation for sustainable science. At the start of 2011, there were an estimated 1895 repositories available for data or literature deposition across all domains, according to numbers provided by the global Directory of Open Access Repositories (OpenDOAR), while in 2022 that figure now exceeds 5800.[13]

In this section we present a general overview of the repository landscape by domain, distinguishing between data and literature repositories. There are also general-purpose repositories like Zenodo[14], provided by OpenAIRE and CERN, which can be limited to specific types of work, subjects, or institutions. The overview is based on available sources such as FAIRSharing.org, re3data, OpenDOAR and data stemming from the MOAP study (Monitoring the Open Access Policy of Horizon 2020) that was commissioned by the European Commission (Manola, Papageorgiou, Grypari, & Lempesis, 2021a and 2021b).

2.1 Data repositories

When it comes to a definition of research data there is no consensus among different domains and contexts. STEM[15] researchers tend to consider all text formats as literature, whereas SSH[16] researchers often see text formats as data. Data may have different natures: they can derive, for example, from observations, numerical simulations, or experimental activities; they can come in different formats such as numeric, visual, audio, text, digital or non-digital. In some contexts, code is presented as data, whereas in others it is considered as a separate research digital object, with specific workflows and collaborative environments for its development (GitHub[17], GitLab[18]), and for its preservation (Software Heritage[19]). Some data repositories allow software to be deposited as one of their content types, and this is also the case for some literature repositories.

[11] https://www.forbes.com/sites/ciocentral/2013/01/15/big-data-get-ready-for-the-2013-big-bang/

[12] https://insidebigdata.com/2017/02/16/the-exponential-growth-of-data/

[13] https://v2.sherpa.ac.uk/view/repository_visualisations/1.html

[14] https://zenodo.org/

[15] Science, Technology, Engineering, and Mathematics

[16] Social Sciences and Humanities

[17] https://github.com/

[18] https://about.gitlab.com/

[19] https://www.softwareheritage.org/

Diverse definitions exist for research data. In the context of this study, we will refer to research data as defined by the Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020 (European Commission, 2017):

Data refers to information, in particular facts or numbers, collected to be examined and considered as a basis for reasoning, discussion, or calculation.

To complete the general overview of data repositories presented in the following, we made use of two internationally recognised registries - FAIRsharing and re3data. These resources, as further explained in the Methodology section, are overlapping yet complementary in terms of domain coverage and curation. The FAIRsharing registry is manually curated and has an historical focus on the Life Sciences, while re3data is not curated and is less updated, but has a larger corpus of institutional repositories.

Observant readers will note that this overview has a larger Life Sciences section, compared to the other domains. This is reflective not only of the resources used to perform the landscape analysis, but also the level of development, investment, and repository proliferation in the different ERC domains.

Life Sciences

Life Sciences (LS) repositories are aided by a long track record of sustained funding, thanks to the growth of genomics in the early 2000s. This led to the generation of database schemata and ontologies that laid the foundations for good data management in the Life Sciences. In addition, there are a number of well-funded European Research Infrastructures and related Clusters, such as ELIXIR[20], BBMRI[21] and EOSC-Life[22] that not only support repositories financially but also encourage and drive consistency and interoperability. This has led to a number of ‘gold standard’ resources for biological data. These can be found clustered at the EMBL-EBI[23] in Hinxton, UK (Ensembl[24], UniProt[25], ChEMBL[26]) and the Swiss Institute of Bioinformatics (NeXtProt[27], IntAct[28]), alongside other national efforts such as the UK BioBank[29]. Interestingly, many of these ‘gold standard’ repositories do not provide their metadata in an open, fully licenced manner and therefore cannot be considered (as it will become clear in the next chapters) as able to facilitate compliance with the HE MGA. We hope this will change in the near future as awareness of the importance of supporting researchers to comply with funding body mandates grows.

Indicative of the dominance of the Life Sciences in data repositories, a 2017 study of 1381 data repositories registered in re3data found 50% include Life Sciences research in their scope (Kindling, Pampel, van de Sandt et al., 2017).

[20] https://elixir-europe.org/
[21] https://www.bbmri-eric.eu/
[22] https://www.eosc-life.eu/
[23] https://www.ebi.ac.uk/
[24] https://www.ensembl.org/index.html
[25] https://www.uniprot.org/
[26] https://www.ebi.ac.uk/chembl/
[27] https://www.nextprot.org/
[28] https://www.ebi.ac.uk/intact/
[29] https://www.ukbiobank.ac.uk/

Physical Sciences and Engineering

There are over 400 domain-specific repositories that cover Physical Sciences and Engineering (PE) (FAIRsharing; accessed May 2022). A large proportion of these are US-based and focus on geology and astronomy. Some are 'dead' repositories, i.e., they were active repositories that were associated with a particular project or initiative and are now frozen dataset archives; the datasets can be accessed, but no new datasets can be deposited. Within the PE domain, the repositories covering the ERC Panel PE10: Earth System Science are worth mentioning as, on the whole, they exhibit better metadata schemata and a greater consideration of good research data management than those covering other panels of this domain. As we will explain further in the Analysis section, the Earth System researchers often work on large amounts of data that are most valuable if they are aggregated in long time series or collected over large geographical areas; these data therefore need to be managed accurately, thus precipitating the need for a robust disciplinary data repository.

Social Sciences and Humanities

Around 27% of data repositories included in re3data include Social Sciences and Humanities (SH) research in their scope (Kindling, Pampel, van de Sandt et al., 2017). These repositories are spread across the globe, but concentrations can be found in the USA and Europe. Many of these repositories are provided by national or supranational funders and are available to all. In Europe in particular, many are also maintained by transnational consortia, focusing on a particular discipline (e.g., literature, art, politics). An example of a national-level repository within this domain is the Swiss National Data and Service Center for the Humanities (DaSCH) that is funded by the Swiss National Science Foundation (SNSF)[30]. In addition to providing access and long-term preservation of humanities data DaSCH also coordinates DARIAH's (Digital Research Infrastructure for the Arts and Humanities) activities in Switzerland. Another example of a repository relevant to this domain active at the international level is *Phonogrammarchiv*[31], which focuses on providing access and long-term archival for audio-visual research recordings from all disciplines and regions of the world. The repository is associated with the Austrian Academy of Sciences and acts as a CLARIN Knowledge Center since 2015, giving expertise and help free of charge for anyone in the world in the areas of archiving, digitisation, long-term preservation, and re-recording of all kinds of audio-visual material.

Domain-agnostic repositories

Many institutional or national repositories are domain-agnostic, with acceptance criteria relating to membership of an institution or research infrastructure, or location in a specific country. re3data contains metadata records for over 700 institutional repositories. They are extremely diverse in schema, metadata, and quality, but themes can be found. Many institutional repositories in the UK are based on commercial software, such as figshare[32].

[30] https://www.dasch.swiss

[31] https://www.oeaw.ac.at/en/phonogrammarchiv

[32] https://figshare.com/

2.2 Literature repositories

In the context of this study, we refer to 'literature' as:

"the product of research portrayed in a textual form, that can be instantiated as journal articles, conference proceedings, preprints, book chapters and long-text publications such as monographs."

This is in line with Öchsner (2013).

While most literature repositories focus on research outputs from an institution (i.e., are institutional repositories), some large-scale disciplinary repositories stand out, in particular arXiv and Europe PubMed Central (Pinfield et al., 2014). Here, we describe key characteristics of these literature repositories relative to the ERC domains. Although they have a leading position in their domains, other subject-specific repositories to disseminate research articles exist. However, both our analysis of the prevalence of ERC-funded research provided by open access repositories (see the sub-section on Literature repository usage by ERC grantees below) and our candidate repository identification for the inventory (see the Methodology section) support previous findings (Pinfield et al., 2014) that these repositories have a more narrow disciplinary scope and, in turn, are of much lesser size.

We have included preprint servers in this overview, as they cover almost every domain and specific discipline (Hoy, 2020). Preprint servers are a key resource for ERC grantees as they facilitate one of the key Open Science practices defined in the HE MGA: the early and open sharing of research. Preprint servers have gained attention in recent years as more and more publishers have allowed the deposition of submitted author manuscripts.

In some cases, depending on the publisher's policies, after peer-review and publication authors may replace the preprint with the AAM (Author Accepted Manuscript) with no embargo. For example, Elsevier's Open Science Policy related to Article Sharing[33] states that:

"[...]Authors can share their accepted manuscript: Immediately by updating a preprint in arXiv or RePEc with the accepted manuscript".

This kind of policy allows authors to share their accepted manuscript immediately at no cost, and therefore preprint servers in combination with publishers' open access policies have opened up a path for grantees to comply with the HE MGA obligations in terms of Open Science.

Life Sciences

In 2009, Europe PubMed Central[34] (Europe PMC, formerly UKPMC) was launched at the EMBL-EBI in Hinxton. Europe PMC ingests all content from the US Life Sciences literature database PubMed Central and extends its index with other literature and patent sources. To date, the Europe PMC open access subset comprises more than 4.8 million articles available under a Creative Commons licence. More than 700,000 of the articles indexed in PubMed in

[33] https://www.elsevier.com/about/policies/sharing

[34] https://europepmc.org/

2021 were open access in Europe PMC, representing a share of 45% of the articles published during that year.[35]

Full texts in Europe PMC mainly stem from a publisher deposition programme initiated by PubMed Central (PMC)[36], which is maintained by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM) at the US National Health Institutes (NIH). Before inclusion in this programme, a journal has to undergo a thorough evaluation process, which assesses disciplinary scope, scientific quality and technical eligibility.[37] An analysis of the PMC journal list[38], downloaded in November 2022, shows that more than 3,900 Life Sciences journals have contributed to PMC to date. However, different participation levels in terms of issue coverage, embargo period, and access and re-use exist. Most articles in PMC were also made openly available through the publisher's website. At the time of writing, 1,907 journals fully participate in the PMC deposition programme, while 1,605 journals only share some articles, which were published under the hybrid open access business model. 383 journals provide free-to-read articles, prohibiting sharing and re-use without permission. In terms of embargo, the majority of journals make articles openly available without delay, while other publishers release articles after a period of 1 – 36 months. However, most journals with embargo periods make articles available within 12 months or less.

In addition to publisher depositions, authors can also submit manuscripts directly through the Europe PMC plus submission system, if a manuscript is already accepted for journal publication. Furthermore, the underlying research must be funded by a partnering research funder. At the time of writing, Europe PMC supports 37 research funders, including the European Research Council (ERC).

Overall, around 47,000 manuscripts are made open access through the Europe PMC plus submission system. For 81% of manuscripts in Europe PMC, preprint versions are also provided by preprint servers, which have become prevalent to disseminate biomedical research findings in the wake of the COVID-19 health crisis (Fraser et al., 2021). Consequently, PMC and Europe PMC began indexing various preprint servers and link preprints to PubMed-indexed journal articles (Ferguson et al., 2020). At the time of writing, Europe PMC contains information on more than 500,000 preprints. Most preprints in Europe PMC are from Research Square[39] (Springer Nature), bioRxiv[40] and medRxiv[41] (both maintained by the Cold Spring Harbor Laboratory (CSHL), US).[42]

Physical Sciences and Engineering

In 1991, the increase in computer network capacity and the availability of the open-source typesetting software TeX enabled researchers in the high-energy physics community to set up

[35] https://europepmc.org/downloads/openaccess
[36] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[37] https://www.ncbi.nlm.nih.gov/pmc/pub/journalselect/
[38] https://europepmc.org/journalList?titles=current&search=journals&journals=collections&journalid
[39] https://www.researchsquare.com/
[40] https://www.biorxiv.org/
[41] https://www.medrxiv.org/
[42] https://europepmc.org/Preprints

the eprint-archive arXiv[43] (Ginsparg, 1994). Since then, the disciplinary repository has broadened in scope to other fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics, although its coverage varies across disciplines (Larivière et al., 2014).

To date, arXiv comprises more than 2,100,000 open access works. In terms of coverage, Gentil-Beccot et al. (2009) determined that in the 2000s, between 90% and 100% of articles published in five main peer-reviewed high energy physics (HEP) journals were also provided by arXiv in some version. They also found that most HEP researchers point to arXiv as their main source to access research. However, in other research fields, compared to journal content indexed in the Web of Science arXiv's coverage was considerably lower (Larivière et al., 2014).

arXiv co-initiated important standardisation activities in the repository landscape. Most prominently, the Open Archive Initiative has promoted web standards to ensure that open access literature is discoverable, preserved and exchangeable (Van de Sompel & Lagoze, 2000). The resulting OAI-PMH standard to exchange repository metadata has been widely adopted. For example, repositories make use of it to provide information about EC-funded research through the OpenAIRE portal.

Social Sciences and Humanities

A prominent subject-specific example in the Social Sciences and Humanities domain is the RePEc (Research Papers in Economics) initiative[44], which focuses on economics research and related fields. Both journals and institutions can set up 'RePEc archives' to disseminate publications, mainly working papers published prior to the publication in a journal. To date, more than 2,000 RePEc archives were registered. The disciplinary repository Munich Personal RePEc Archive[45] allows authors without access to an institutional RePEc archive to disseminate economics research literature. Another example supporting both institutions and authors is the disciplinary repository EconStor maintained by the German ZBW – Leibniz Information Centre for Economics.[46]

In 2010, the OAPEN Library[47], maintained by the not-for-profit organisation OAPEN Foundation based in the Netherlands, was launched as a literature repository for open access books, particularly facilitating dissemination of open access scholarly books in the Humanities (Eve, 2014). At the time of writing, OAPEN hosts more than 22,000 books and 2,800 book chapters from both commercial publishers such as Taylor & Francis or Springer Nature and university presses like Firenze University Press or Amsterdam University Press. ERC recommends deposit in the OAPEN Library. Accordingly, OAPEN provides guidance for ERC funded researchers.[48]

[43] https://arxiv.org/
[44] http://repec.org/
[45] https://mpra.ub.uni-muenchen.de/
[46] https://www.econstor.eu/
[47] https://www.oapen.org/
[48] https://www.oapen.org/topic/erc-european-research-council

Literature repository usage by ERC grantees

To better understand the use of repositories to disseminate ERC-funded research, we drew on open data compiled as part of the MOAP study (Monitoring the Open Access Policy of Horizon 2020) commissioned by the European Commission (Manola, Papageorgiou, Grypari, & Lempesis, 2021a and 2021b). The database itself is based on the OpenAIRE Research Graph (Manghi et al., 2021), which comprises multiple scholarly data sources, including repositories meeting the OpenAIRE guidelines. These guidelines provide repository managers with a standardised mechanism for data exchange including grant information. Because of the networked and standardised reporting within the OpenAIRE network, the OpenAIRE Research Graph is considered to be more comprehensive than research publications reported by EU research grant holders alone (Mugabushaka et al., 2021).

The aim of the MOAP study, which was completed in June 2021, was to present the state of compliance with the Horizon 2020 open access requirements, for both publications and research data. A limiting factor of the MOAP study is that data were gathered already in January 2021, likely not covering all publications reporting H2020-funded research. In particular, the authors noted considerable delays in the overall process of open access reporting in H2020, indicated by a low number of reported research outputs published in 2020 compared to the period 2014-2019. Furthermore, they observed a comparatively low number of publications from H2020 projects in the Social Sciences and Humanities, highlighting large variations by discipline in terms of time between the start of grants and the publication of research arising from projects.

After matching and subsetting the MOAP data to ERC grants, we also found large discrepancies between data and literature repositories in terms of coverage. While we were able to link 58% of ERC grants under H2020 to literature repositories, only 5% could be matched to data repositories, presumably because not all projects participated in the H2020 Open Research Data Pilot (ORDP). Because of this low coverage, we only considered literature repositories in the following MOAP analysis. The objective was to determine literature repository usage measured by the number of publications by ERC domain and panel, highlighting widely used repositories as well as disciplinary differences.

Domain	Total publications	ERC projects	% of Total projects
Life Sciences	10.404	1.479	60.1
Physical Sciences and Engineering	25.987	2.233	62.1
Social Sciences and Humanities	4.512	745	44.6
no match*	198	34	32.4

Table 1: Literature repository coverage by ERC domain (own calculation using MOAP data and internal ERC mapping table of projects to panels and domains). *Note that projects supported by a ‘Synergy’ or ‘Proof-of-Concept’ grant are not associated with ERC panels and could therefore not be matched to an ERC domain. Note that publications may be associated with one or more ERC projects.

Overall, we were able to identify 40,374 peer-reviewed publications linked to ERC projects under H2020. Table 1 provides a breakdown by ERC domain. Both in terms of publications and projects, ERC projects in the domain of the Physical Sciences and Engineering (25,987 publications across 2,233 out of 3,598 projects funded by the ERC) were best represented in our sample, followed by the Life Sciences (10,404 publications across 1,479 of 2,459 projects) and the Social Sciences and Humanities (4,512 publications across 745 of projects). The lower coverage of ERC-funded projects in the Social Sciences and Humanities domain observed in our sample may reflect that publications in the ERC Social Sciences and Humanities domain are published after the end of the project to a greater extent than in the Life Sciences and Physical Sciences and Engineering domains.

Overall, we were able to identify 464 repositories covering peer-reviewed research literature linked to ERC projects. Figure 1 shows the Top 10 repositories for each ERC domain in terms of publications arising from ERC projects in H2020. It highlights the prominent role of Europe PubMed Central and arXiv among ERC grantees: around 75% of publications in the Life Sciences domain were available in Europe PubMed Central (7,881 research publications), while arXiv recorded more than 50% of articles in the Physical Sciences and Engineering domain (13,762 research publications). Comparing the three ERC domains highlights different degrees of concentration. While in the Life Sciences domain more than 75% of publications could be attributed to one repository, Europe PMC, the distribution of deposits in the Social Sciences and Humanities is much less concentrated.

The dominant position of Europe PMC in the Life Sciences domain is not surprising owing to its publisher deposit programme and manuscript submission system, which allows authors to link their work to ERC funding. In the Life Sciences domain, Europe PMC is followed by repositories belonging to the French Hyper Article en Ligne (HAL) network[49], which primarily serves authors from French research institutions. As an external link provider, HAL also interlinks its content with Europe PMC. HAL is followed by another national aggregator, the Dutch NARCIS portal[50] and MPG.Pure[51], the institutional repository of the German Max-Planck Society. Other institutional repositories in the Top 10 of the Life Sciences domain were the Oxford University Research Archive (ORA)[52], the Leuven Institutional Repository and Information Archiving System (LIRIAS)[53], the University of Cambridge Repository Apollo[54], DIGITAL.CSIC[55], the institutional repository of the Spanish National Research Council, and ZORA (Zurich Open Repository and Archive)[56] from the Swiss University of Zurich. Many of these institutional repositories were also relatively well represented in the other two ERC domains, highlighting the multi-disciplinary scope of institutional repositories. The preprint server bioRxiv also ranked among the Top 10 in the Life Science domain.

[49] https://hal.archives-ouvertes.fr/

[50] https://www.narcis.nl/

[51] https://pure.mpg.de/

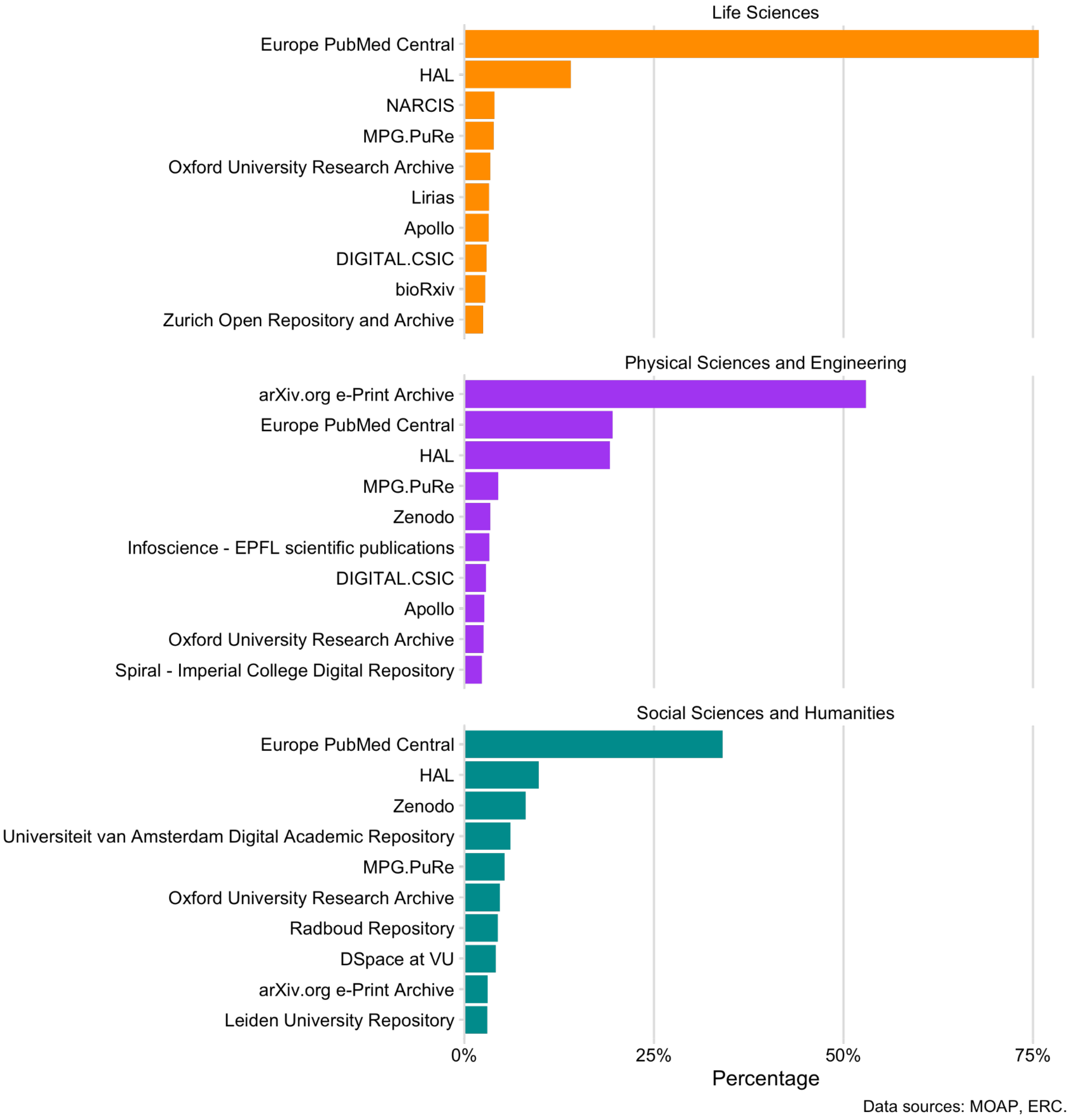
[52] https://ora.ox.ac.uk/

[53] https://www.kuleuven.be/english/research/scholcomm/lirias

[54] https://www.repository.cam.ac.uk/

[55] https://digital.csic.es/

[56] https://www.zora.uzh.ch/



Data sources: MOAP, ERC.

Figure 1: Top 10 widely used repositories for each ERC domain. Note that a publication can be deposited in more than one repository.

Surprisingly, Europe PMC, with its focus on biomedical literature is also well represented in the Physical Sciences and Engineering and Social Sciences and Humanities domains, suggesting disciplinary links to the Life Sciences. A closer look into the distribution across ERC panels[57] reveals that publishing in Life Sciences-related journals indexed by Europe PMC was predominant in PE04: Physical and Analytical Chemical Sciences and PE05: Synthetic Chemistry and Materials. In the Social Sciences and Humanities domain, Europe PMC accounted for more than half of research literature originating from ERC projects in the panel SH04: The Human Mind and Its Complexity, due to the association of the field of

[57] An interactive supplement provides insights into the ERC usage of repositories by panel. https://erc-repo-usage-h2020.netlify.app/

Neuroscience with both the Life Sciences and psychology. Like in the Life Science domain, institutional repositories were well represented in the Physical Sciences and Engineering (Infoscience from the Swiss École Polytechnique Fédérale de Lausanne (EPFL) and Spiral form the British Imperial College) and Social Sciences and Humanities domain. Interestingly, there are three Dutch institutional repositories among the Top 10 in the Social Sciences and Humanities, belonging to the Universiteit van Amsterdam, Vrije Universiteit Amsterdam (VU) and Radboud Universiteit.

We have also investigated the different levels of compatibility with the OpenAIRE guidelines by matching MOAP data to the OpenAIRE Research Graph data source table (Manghi et al., 2021). The objective was to obtain an initial estimate of literature repository alignment with the Open Science Horizon Europe MGA requirements to inform the inventory design. Table 2 presents the different levels of OpenAIRE compatibility, indicating for each of them the number and percentage of repositories and deposits of research literature. The table is sorted by the most recent version of the OpenAIRE guidelines[58] in descending order. Since the version OpenAIRE 2.0, the guidelines describe an automated mechanism for metadata exchange to meet open access funder mandates. Interestingly, the table highlights that 45% of literature items could be attributed to repositories that are tagged as 'collected from a compatible aggregator'. OpenAIRE harvests metadata from these repositories, to which Europe PMC belongs. However, we were not able to find documentation of the underlying mechanism for metadata transfer. Also, a large proportion of publications were deposited in repositories that do not share grant information ('OpenAIRE Basic (DRIVER OA)'). arXiv is a prominent example in this regard.

OpenAIRE compatibility	Repositories	% of Total	Deposits	% of Total
OpenAIRE PubRepos v4.0*	18	3.9	1625	4
OpenAIRE 3.0 (OA, funding)*	187	40.3	18077	44.2
OpenAIRE 2.0+ (DRIVER OA, EC funding)*	47	10.1	4311	10.5
OpenAIRE Basic (DRIVER OA)	134	28.9	19912	48.6
collected from a compatible aggregator	63	13.6	18575	45.4
not available	16	3.4	1043	2.5

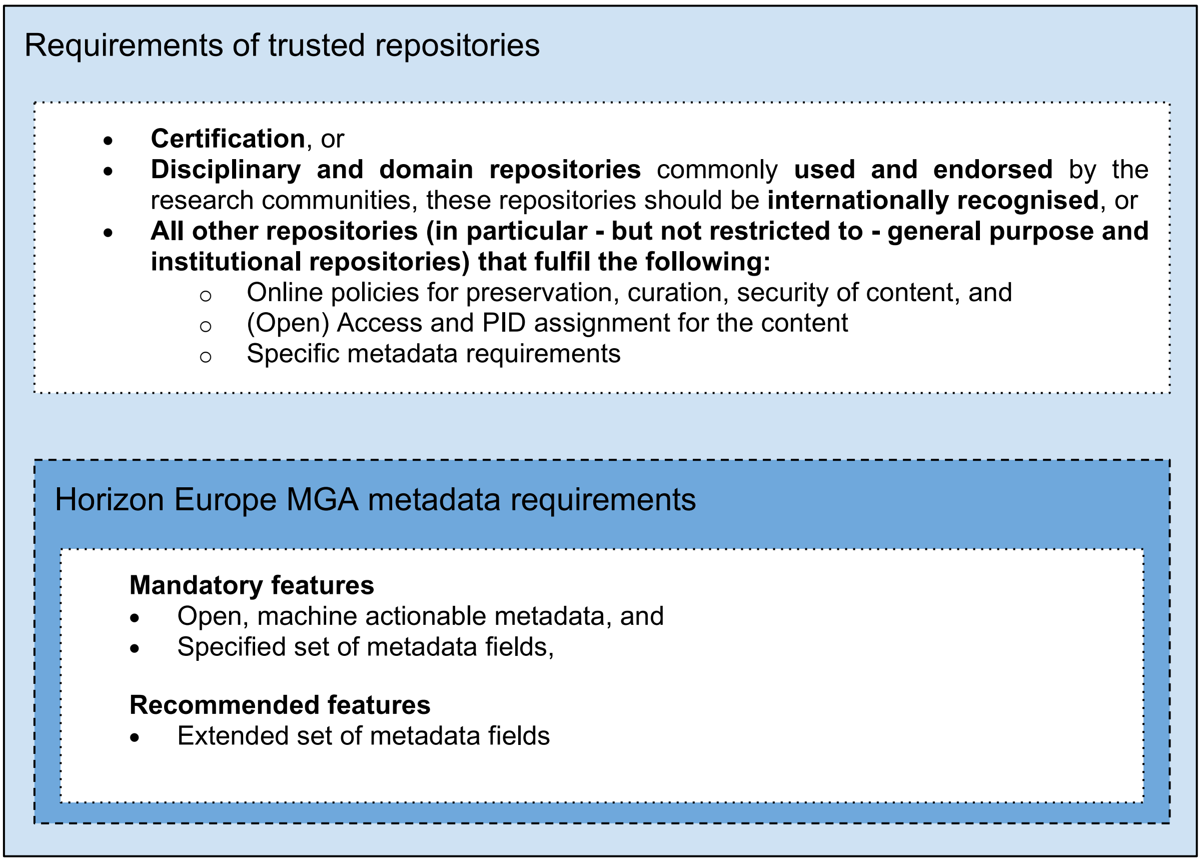
Table 2: Different levels of compatibility with the OpenAIRE guidelines for literature repositories that have been used by ERC grantees to deposit the publications resulting from their projects. *Repositories meeting Horizon 2020 Open Access Requirements. Note that a publication can be deposited in more than one repository.

[58] https://guidelines.openaire.eu/en/latest/

3. Concepts

In this section we will present the basic concepts on which the study was based.

Figure 2 presents an overview of the requirements for trusted repositories as set out in the HE AGA, and of the HE MGA in terms of metadata.



Requirements of trusted repositories

- **Certification**, or
- **Disciplinary and domain repositories** commonly **used and endorsed** by the research communities, these repositories should be **internationally recognised**, or
- **All other repositories (in particular - but not restricted to - general purpose and institutional repositories) that fulfil the following:**
  - Online policies for preservation, curation, security of content, and
  - (Open) Access and PID assignment for the content
  - Specific metadata requirements

Horizon Europe MGA metadata requirements

Mandatory features

- Open, machine actionable metadata, and
- Specified set of metadata fields,

Recommended features

- Extended set of metadata fields

Figure 2: Overview of requirements for trusted repositories defined in the HE AGA and HE MGA metadata requirements.

The HE MGA Art. 17 requires grantees to deposit their outputs in trusted repositories. In particular, concerning peer-reviewed scientific publications relating to their result, they must ensure that:

“...at the latest at the time of publication, a machine-readable electronic copy of the published version or the final peer-reviewed manuscript accepted for publication, is deposited in a trusted repository for scientific publications”

The grantees also must:

“...as soon as possible and within the deadlines set out in the DMP, deposit the data in a trusted repository; if required in the call conditions, this repository must be federated in the EOSC in compliance with EOSC requirements”

On top of the mandatory use of trusted repositories, the HE MGA includes a set of metadata requirements, linked to the Open Science mandate.

A selected set of 220 repositories has been included in this study to be tested towards their readiness to facilitate ERC grantees to comply with the HE MGA mandates.

3.1 Trusted repositories

The HE MGA calls for grantees to utilise repositories that fulfil a set of criteria related to trustiness. A repository can be considered trusted, as defined in the HE AGA and further clarified by the ERCEA, if it is certified, or in the case of disciplinary and domain repositories, if it is commonly used and endorsed by an appropriate research community and recognised internationally, or for general-purpose or institutional repositories, or repositories that do not fall into the previous categories, if they meet certain more detailed requirements.

Certification

The HE AGA considers certified repositories that have received the CoreTrustSeal, the Nestor Seal DIN31644, the ISO16363 certification, or similar, to be automatically trusted. Corrado (2019) provides a summary of the histories and general approaches and requirements of these three certifications. However, as described in the introduction to this report, only a small fraction of existing repositories have such certifications.

Community endorsement and international recognition for disciplinary and domain repositories

The exact definition of what counts as endorsement by a research community cannot be precisely derived solely from the text provided in the HE AGA (page 155):

“...disciplinary and domain repositories commonly used and endorsed by the research communities. Such repositories should be recognised internationally.”

There is no single standardised system for organisations or individuals to codify and express commitment to and endorsement of specific repositories. For example, endorsement can be seen to happen through use, so that repositories popular among researchers can be considered endorsed by their research community.

Moreover, to be considered trusted, community endorsed disciplinary or domain-specific repositories need to be internationally recognised. For the purpose of this study, we define this category as those repositories that are acknowledged by the international research community, through badges or accreditation. National repositories may fall into this category if they are internationally recognised as the repository that serves the national research community. For the purpose of this study, we assumed that inclusion in a formal international registry (such as OpenDOAR, re3data or FAIRsharing) can be accepted as a proxy for recognition.

It is important to note that we have categorised repositories as trusted when there has been sufficient data available to do so. Otherwise, the category ‘Data inconclusive’ has been used.

Features required for repositories that do not fall under previous categories

When a repository is neither certified nor endorsed by a specific community (and therefore does not fall in the previous categories), the repository needs to fulfil some essential characteristics listed in the HE AGA to be considered trusted. These are listed in the subsections below and were also presented as part of Figure 2. Such characteristics are linked to preservation, curation and security of content, Open Access and PID assignment, and specific metadata requirements.

Preservation, curation, and security of the content

For scholarly journals, books, and other types of published content, publishers can enrol into one or several preservation services (e.g., Portico[59], CLOCKSS[60]) that ensure the long-term archiving of content. The existence of such services makes it easier to find out what kind of standardised preservation measures are implemented, and more fundamentally, if a content provider is enrolled or not since this information is often made public through the service provider. For repository operators there are no such widely used assured international networks in use, making it much harder to gauge what measures a repository has taken towards preservation, curation, and security of content. The definition of trusted repository in the HE AGA mentions that repositories should "have specific provisions in place and offer explicit information online about their policies, which define their services"; for the purposes of this study having a public policy in place is considered sufficient for a repository to fulfil this criterion.

(Open) access and PID assignment for the content

Repositories are asked to "provide broad, equitable and ideally open access to content free at the point of use, as appropriate, and respect applicable legal and ethical limitations. They assign persistent unique identifiers to contents (e.g., DOIs, handles, etc.), such that the contents (publications, data and other research outputs) are unequivocally referenced and thus citeable."

For this reason, we identified as features to be assessed:

- the capacity of the repository to assign a PID to the content
- potential restrictions for accessing repository content

To support this process, we referred to the definition given by re3data, as this is already established in the repository community (see Box 1).

In this study, we consider only repositories that implement an Open or Restricted access to their contents as Closed repositories (per re3data definition in Box 1) cannot meet the requirements set out in the HE MGA and HE AGA.

[59] Portico, a community-supported preservation archive that safeguards access to e-journals, e-books, and digital collections https://www.portico.org/

[60] CLOCKSS (Controlled LOCKSS), a sustainable dark archive to ensure the long-term survival of digital scholarly content https://clockss.org/

BOX 1: re3data definition of Access to the repository contents

Access to the repository content refers to the general content of the repository and not to single access rights for specific records in the repository. This access can be:

- Open: external users can access the content of the repositories with no barriers
- Restricted: external users can overcome access barriers (e.g.: the repository requires authentication/authorisation to access the content)
- Closed: external users cannot overcome access barriers (e.g.: the repository content is only available to internal users managing the infrastructure)

Metadata requirements

A set of requirements for metadata is also defined in the HE AGA as an essential characteristic of trusted repositories. In particular, after discussing and further clarifying the details with ERCEA, we addressed them dividing them in two encompassing features:

- Metadata should contain information about licensing
- Metadata should be machine-actionable and standardised

They will be elaborated in the following section. Requirements related to the FAIRness of the repository were not included in the study as many tools are already available for ERC grantees to assess this. Other requirements that could be subject to interpretation were also not included in the study, such as the fact that repositories “ensure that contents are accompanied by metadata sufficiently detailed and of sufficiently high quality to enable discovery, reuse and citation and contain information about provenance” (European Commission, 2021, page 155).

3.2 Horizon Europe MGA repository requirements

Content access and licensing

The HE MGA outlines the following requirements concerning content access and licensing:

"Immediate open access is provided to the deposited publication via the repository, under the latest available version of the Creative Commons Attribution International Public Licence (CC BY) or a licence with equivalent rights; for monographs and other long-text formats, the licence may exclude commercial uses and derivative works (e.g. CC BY-NC, CC BY-ND)”

“as soon as possible and within the deadlines set out in the DMP, ensure open access — via the repository — to the deposited data, under the latest available version of the Creative Commons Attribution International Public licence (CC BY) or Creative Commons Public Domain Dedication (CC 0) or a licence with equivalent rights”

For fulfilment of these criteria, we consider the repository to be required to provide a field in the metadata in which the licence of the deposited item can be indicated.

Metadata access and licensing

The Horizon Europe MGA outlines the following requirements for repository deposit concerning metadata access and licensing for publications:

“Metadata of deposited publications must be open under a Creative Common Public Domain Dedication (CC 0) or equivalent, in line with the FAIR principles (in particular machine actionable) and provide information at least about the following: publication (author(s), title, date of publication, publication venue); Horizon Europe or Euratom funding; grant project name, acronym and number; licensing terms; persistent identifiers for the publication, the authors involved in the action and, if possible, for their organisations and the grant. Where applicable, the metadata must include persistent identifiers for any research output or any other tools and instruments needed to validate the conclusions of the publication.”

And for data:

"Metadata of deposited data must be open under a Creative Common Public Domain Dedication (CC 0) or equivalent (to the extent legitimate interests or constraints are safeguarded), in line with the FAIR principles (in particular machine-actionable) and provide information at least about the following: datasets (description, date of deposit, author(s), venue and embargo); Horizon Europe or Euratom funding; grant project name, acronym and number; licensing terms; persistent identifiers for the dataset, the authors involved in the action, and, if possible, for their organisations and the grant. Where applicable, the metadata must include persistent identifiers for related publications and other research outputs."

For fulfilment of these criteria, we consider the repository to be required to provide metadata related to each digital object via a public domain dedication such as CC0, Public Domain or equivalent. However, as it is further shown in the Analysis section many repositories do not readily provide this information in a clear human or machine-readable manner.

4. Methodology

In this section we describe the resources we used to identify candidate repositories for the study, the selection criteria we used to reach a suitable number of trusted repositories to be assessed, and the resulting inventory of identified trusted repositories. This section also depicts the data collection strategy.

The core output of this work, together with this report, is the inventory to be used by ERC grantees. It will enable them to make informed decisions as to which repositories to use for their data or publication in order to comply with the HE MGA requirements. We describe how the final inventory of repositories was designed and created as an output of the assessment work. We have made an effort to detail the methodology employed such that it may be reproduced and improved upon by the community. We see the inventory as a first release and snapshot of the current state and expect it to evolve over time.

4.1 Repositories selection methodology

Starting from the results of the general overview of the repository landscape (which have been presented in Section 2) we selected a first set of candidate repositories to be representative of the landscape and of the three ERC domains (PE, LS, SH).

Our goal was to have a list of repositories to be included in the study for which we could manage the information collection, which was carried out as a combination of survey and manual curation (we will cover survey creation in later sections). We set our goal to be less than 500 repositories for the first candidate selection and between 200 and 300 repositories to be assessed in the study.

Figure 3 depicts the four phases of the selection process (central arrows), which are presented together with the number of repositories included in each phase (top circles) and the actions carried out to proceed from one phase to the other (bottom boxes).

The selection process is briefly summarised below:

1. **General overview of the repository landscape** - this includes all repositories in the public sources we consulted for results in Section 2;
2. **Candidate repositories for the study** - a rough selection aimed to obtain a first list of repositories to be tested against inclusion criteria; this list was mainly designed to have a good coverage of specific repositories linked to each ERC panel, and to have a representation of those mostly used by ERC grantees in the past for deposit;
3. **Repositories included in the study** - the repositories that have been included in the study and for which we collected detailed information about their policies, metadata and coverage; these repositories were tested towards their readiness to facilitate the ERC grantees to comply with the HE MGA mandates;
4. **Inventory of identified trusted repositories** - repositories that have been identified to be trusted.

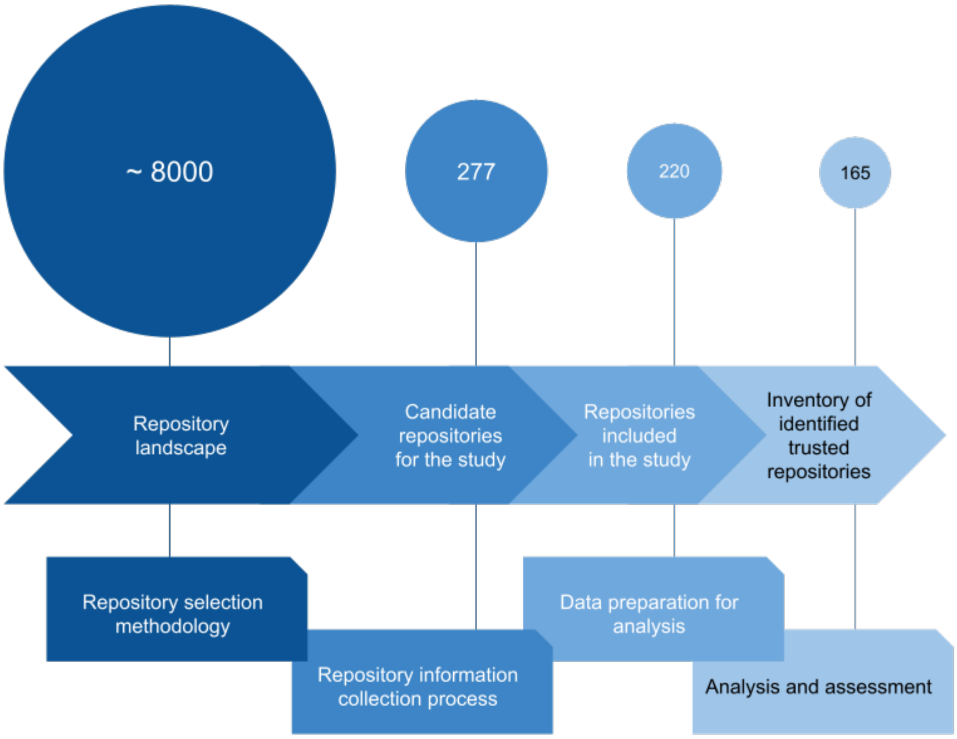

~ 8000
277
220
165
Repository landscape
Candidate repositories for the study
Repositories included in the study
Inventory of identified trusted repositories
Repository selection methodology
Repository information collection process
Data preparation for analysis
Analysis and assessment

Figure 3: Diagram depicting the four phases of the selection process (central arrows), which are presented together with the dimension of the set of repositories to be handled in the various phases (top circles) and the actions carried out to proceed from one phase to the other (bottom boxes).

To select the list of candidate repositories, different sources of information were used. It is important to highlight that, before the repository information collection process, little information was available for each repository; moreover, the type of information available was linked to the source. For example, data on the repository content type (data, literature, catch all), and the openness of the repository in terms of upload or contents, were not trivial to obtain, as for some of the sources this information was not updated or not provided in a machine actionable way. We present in the following the process that led to the list of candidate repositories for the study.

Extraction of repositories from the ERC Information document

One of the sources we used to extract candidate repositories for the study was derived combining the repositories recommended by the ERC and those mentioned in the ERC Scientific Council's information document 'Open Research Data and Data Management Plans' (European Research Council Scientific Council, 2022).

From the ERC information document, a list of repositories was extracted, together with a set of related information for each repository (repository name, URL, ERC domain). The information document explicitly mentions a set of repositories divided by ERC domains together with domain-specific research infrastructures (ELIXIR, CESSDA, EPOS, etc.) whose websites were searched for repositories to be included in the list. This latter work was not so trivial as research infrastructures tend to present the resources in different ways on their website and it is not always straightforward to find and extract information about specific services such as repositories.

A list of 64 data, literature and catch-all repositories, both general and domain-specific was extracted. These repositories were fed into the selection process for data and literature repositories as explained in the sub-sections below.

Data repository sources and selection

Focusing on quality, we identified FAIRsharing.org, re3data.org and the CTS website as three valuable sources of information for data repositories. It is worth noting that, even if we are referring to these as sources of information for data repositories, they are not limited to repositories that can accept only data, but sometimes include repositories hosting also other types of research items, such as literature, or software, or even catch all repositories.

Through the use of the open and accessible APIs on re3data and FAIRsharing, we were able to extract metadata for over 2,000 data repositories. The teams behind these resources were very responsive when we could not extract the metadata via their APIs.

We discovered a number of quality issues with the publicly available metadata, some of which are outlined in the following. We found both re3data and FAIRsharing to contain a number of factual errors. For example, in re3data some institutes, organisations and standards are listed as repositories. These are easily spotted by domain experts but are missed via computational methods. The FAIRsharing dataset was much more accurate in this respect and, consequently, we have based our initial repository set on the FAIRsharing list of 1,142 repositories.

Due to FAIRsharing's historical focus on the Life Sciences, we have also used a number of other resources to obtain a set of data repositories that could be representative of all the ERC domains. These included the ERC Scientific Council's information document 'Open Research Data and Data Management Plans', mentioned in the previous section, including a mix of data, literature, and catch all repositories. Another resource was the CTS website, from where we extracted 121 repositories. Where available, our metadata were crossed-referenced and checked with metadata from re3data, the Research Data Alliance[61], and the EOSC Marketplace[62].

This combined list of repositories, initially about 2,000, was then cleaned. Entries were deduplicated, errors were removed, and all repositories were reviewed. Errors included repositories appearing more than once both as a repository and as the organisation behind the repository, some standards being mistaken as repositories in re3data, some databases and knowledge bases being classified as data repositories as they have some user curation (e.g., FlyBase[63]), and some institutional repositories being closed archives rather than repositories for data deposition. After this initial clean, we had a list of approximately 500 repositories. Those repositories that have been endorsed by CTS, or European infrastructures (e.g., ELIXIR or CLARIN) were taken forward, while those repositories that were smaller, or perhaps parallel to existing, larger, better-adopted repositories, were removed. We then whittled the repositories down further by removing those repositories that didn't have a declared sustained funding stream. We reviewed the list to remove repositories that were considered too niche or too far geographically from where ERC grantees would work. An

[61] https://www.rd-alliance.org/
[62] https://marketplace.eosc-portal.eu/
[63] https://flybase.org/

example here is the Consortium of Pacific Herberia[64]. Finally, we then assessed coverage of the ERC domains and panels, noting that some panels, such as PE2: Fundamental Constituents of Matter, were only represented by domain-agnostic repositories. We then went back to FAIRsharing and re3data to identify repositories that we may have missed or misclassified during our triage process (such as HepData[65], from CERN). This resulted in a final list of 177 repositories.

Literature repository sources and selection

We used OpenDOAR as a main source to investigate literature repositories. As for the data repository sources, we highlight that OpenDOAR is not only listing repositories dedicated to literature only, but includes also those accepting other kinds of outputs, as well as catch-all repositories. Through the OpenDOAR API, we obtained information on 5857 repositories. As a first step, we created a subset of repositories tagged as disciplinary repositories. Our sample consisted of 366 disciplinary repositories, confirming that OpenDOAR mainly comprises institutional repositories (Pinfield et al., 2014). Next, we did a manual eligibility check and removed repositories that were out of scope for our study. We manually examined whether repositories allow self-archiving of current research publications from specific subjects and removed from the list those containing digital collections of historical documents or learning materials. While doing so, we also validated if these repositories were still active, observing 89 URLs indexed in OpenDOAR that did not lead to the repository, which is in line with recent research about repositories' link availability (Mannocci et al., 2022); 7 repositories had stopped accepting new submissions, but were still online. We also excluded grey literature repositories as well as journal and proceedings platforms. Some were excluded because they collected the output of a single research institution or because they seemed to be hacked and led us to phishing websites.

Overall, we found 60 disciplinary literature repositories meeting our selection criteria. Of those, 19 disciplinary repositories provided open access to literature arising from ERC-funded research according to our MOAP analysis (see 2.2). To acknowledge literature repository usage by ERC grantees, we decided to extend our sample of disciplinary repositories with institutional and aggregating repositories, including only those non-disciplinary repositories that belong to the top 10% in terms of deposits compared with other repositories per ERC panel. In total, the resulting list comprised 120 repositories.

At the end of the selection process, we made sure that all three literature repositories that are recommended by the ERC (Europe PMC, arXiv, and OAPEN library) were included in the study and we merged the list with the one obtained in the previous selection steps. Repositories that appeared more than once were deduplicated. A final combined list of 277 repositories was created as candidates to be included in the final inventory

4.2 Repository information collection process

For the next steps towards creating the inventory, we derived repository requirements that facilitate the grantees' compliance with the HE MGA. These were collected and coded in a standardised way from the HE MGA into a selection of features. The structure of the inventory

[64] https://www.pacificherbaria.org/

[65] https://www.hepdata.net/

was built to accommodate repositories and corresponding features in a spreadsheet where each row identifies a repository, and the columns present the values for the repository features (see 4.3 Building the inventory). While some of these features could be derived in an automatic or semi-automatic way, the information accessible in this way is not comprehensive enough to completely describe the level of readiness of the repositories. Therefore, most of the features had to be manually collected. In order to improve the accuracy and speed of this manual curation, we designed a survey for completion by the repository managers themselves. The survey was distributed to all the 277 candidate repositories and 163 answers were collected. For the remaining repositories, we collected information through identified sources and from the repository websites.

Survey design

Based on the envisaged inventory features, a survey was designed to collect information directly from repository managers. This choice was made for two main reasons:

- Limited time and large number of repositories. The direct collection of data from repository managers allowed us to concentrate our efforts and to include a larger number of repositories in the list.
- Direct involvement of repository managers in the study. The involvement of repository managers allowed us to reduce the errors in the data collection process and also gave us an opportunity to collect comments from the community before the results of the study become public

The survey is available as ANNEX 2 to this report. It was designed to contain mainly multiple-choice questions with a limited number of free text questions. Some of the features that could be automatically retrieved were not included in the survey (e.g., OpenAIRE compliance); these were added in the inventory after the survey had been closed. No personal information was collected through the survey. The repository managers surveyed were informed about the scope of the study and the envisaged publication of its results. A guide was included for each question of the survey to reduce the risk of misinterpretation by the respondents. When relevant, links to external useful information and reference links were provided.

Collection of contact information

E-mail addresses representing literature repository contacts were retrieved by querying OAI-PMH interfaces for information about the repository, where we parsed the repository administrator e-mail address from the adminEmail xml node. For data repositories, the contact information was extracted from re3data and FAIRsharing databases. In cases where no repository administrator contact was provided, we inspected the homepages of the repositories. In a few cases, repositories did not provide a contact e-mail address, but rather a contact form.

In July 2022, 277 repositories were invited via emails and contact forms to fill in the survey. Two reminders were sent, and the survey was closed at the beginning of September 2022. A total of 163 responses were collected from repository managers. For the remaining 112 non-responding repositories, we collected the information through the process described in the following section.

Manual data collection for non-responders

Manual curation is not a trivial act. For data repositories, this process involved the consultation of re3data and FAIRsharing metadata records to look for specific features such as the standards used for metadata, the repository policies, and the existence of some certification. We have been unable to automatically extract all required metadata properties from the metadata records in FAIRsharing and re3data. For literature repositories, we could not use OpenDOAR as the information provided did not fit with our survey. For each repository that was manually curated, we performed a detailed investigation. Often a repository will not provide all the information necessary, particularly in relation to policies, and this can lead to a number of investigative steps. In some cases, we were able to retrieve all the information needed only after trying to upload an item to the repository, in other cases this was not possible as the repository only provided upload via institutional or community login.

For the above-mentioned reasons, manual curation could take up to 20 minutes for a single repository.

During this step the manual curation led also to the exclusion of some of the original candidate repositories. This exclusion was due to several reasons:

- The candidate repository was not open for upload by ERC grantees (e.g., only open for submission of outputs from research funded by other funders or falling under the definition of ‘Closed Data Upload’ given by re3data - see Box 2 below).
- The candidate repository was closed for submission or no longer updated.
- The candidate repository was identified as a database.

Regarding the definition of access to upload in the repository, we relied on the one given by re3data in its F.A.Q. section (Box 2). When repositories were found to be closed to upload, they were removed from the list of candidates.

BOX 2: re3data definition of Access to the repository upload

Access to the repository upload refers to the general upload of content to the repository and not to single access rights for specific records in the repository. This access can be:

- Open: external users can upload the content to the repositories with no barriers
- Restricted: external users can overcome access barriers (e.g.: the repository requires authentication/authorisation to upload content)
- Closed: external users cannot overcome access barriers (e.g.: only internal users can upload content to the repository)

Data preparation for analysis

At the end of the repository information collection process, we prepared the data for the analysis. To obtain information about the level of compliance with OpenAIRE, we manually checked the OpenAIRE Explore portal.

Duplicates were deleted or merged, when more than one survey response was given for the same repository. This happened especially for those repositories where a mailing list is provided as contact (we received more than one response for the same repository).

Incomplete responses were deleted when the information provided was insufficient to perform the analysis (e.g.: list of metadata was empty, no information was provided for trusted repository assessment).

Obsolete repositories or repositories that were merged to others and are no longer updated were removed. We also removed repositories that were not relevant for ERC grantees because of scope or closed access to upload.

The manual curation described in the previous section and the data preparation described above led to a final list of 220 repositories (both data and literature) that were further analysed and assessed.

Repository assessment

The assessment was based on the following principles:

- The most machine-actionable path was prioritised. Prioritising clear, simple decisions not only enables consistent, accurate curation, but facilitates future machine actionability of the repository spreadsheet.
- The assessment criteria were defined to be simple, easy to implement and non-controversial.

To assess the compliance with the trusted repository definition, the decision process depicted in Figure 4 was used.

Table 3 presents a summary of the assessment of compliance with the trusted repository definition in the HE AGA, and alignment with the metadata requirements in the HE MGA.

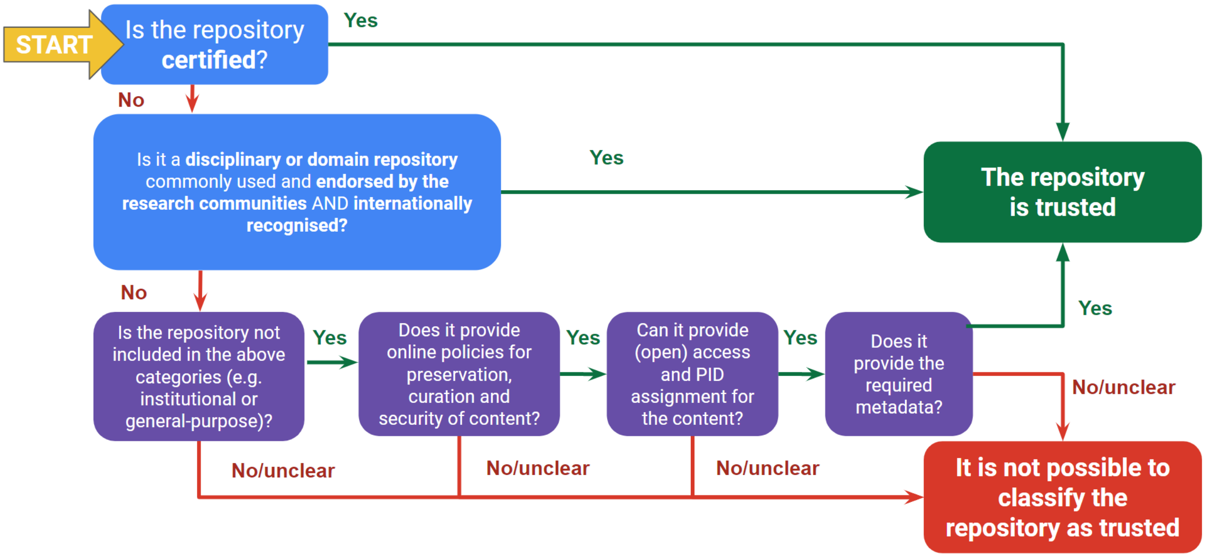

START
Is the repository certified?
Yes
No
Is it a disciplinary or domain repository commonly used and endorsed by the research communities AND internationally recognised?
Yes
The repository is trusted
No
Is the repository not included in the above categories (e.g. institutional or general-purpose)?
Yes
Does it provide online policies for preservation, curation and security of content?
Yes
Can it provide (open) access and PID assignment for the content?
Yes
Does it provide the required metadata?
Yes
No/unclear
No/unclear
No/unclear
No/unclear
It is not possible to classify the repository as trusted

Figure 4: Decision process for trusted repository compliance assessment.

The features were then used to build a spreadsheet where the repository characteristics are described.

Features to be assessed	Classification	Description
Trusted repository features	Compliant	The repository is trusted
	Data inconclusive	Classification not performed
Completeness of metadata information	Basic	The repository includes AT LEAST basic metadata (Author(s), Description, Title, Date of publication).
	Mandatory	The repository includes AT LEAST all the metadata necessary for grantees to provide the mandatory metadata specified in the HE MGA.
	Mandatory and recommended	The repository includes ALL the metadata necessary for grantees to provide both mandatory and recommended metadata specified in the HE MGA.
	Data inconclusive	Classification not performed

Table 3: Repository features to be assessed and related classification options

4.3 Building the inventory

Some basic principles were agreed and formed the base for each further decision:

- The final inventory is for the grantees to use: it should be an easy tool to identify a suitable repository for their outputs.
- The inventory should take the form of a simple spreadsheet with features (columns) and repositories (rows).
- Where possible, controlled vocabularies should be used to describe the specific repository features, building on already existing criteria/best practices/classification.

The features included in the inventory were grouped into the following categories, where the two first ones provide descriptive background information about the repository and the two latter ones relate to trustiness and HE MGA requirements:

- General information about the repository
- Repository usage
- Trusted repository features
- Alignment with the requirements and recommendations related to metadata in the HE MGA

In the following, we describe each set of features and their role in the assessment of alignment with the HE AGA and HE MGA constraints.

It is important to note that the information collected via the survey is self-reported and may contain errors due to possible misunderstanding by the repository managers who answered the survey. For non-responding repositories for which the expert group sought out publicly available information there might also be errors due to the availability and interpretation of this information.

General information about the repository

This set of features includes general information for each repository. It was possible to collect this data automatically from a variety of sources. These metadata did not contribute to the decision for inclusion in the inventory. The set of features is described in Table 4.

Feature	Description
Repository name	The official name of the repository
URL	The direct link to the repository homepage URL

Table 4: Repository features related to general information about the repository

Repository usage

In order to guide the grantee in the selection of the most appropriate repository, specific features have been identified to describe the repository usage. In particular, Table 5 shows repository usage features that were included in the inventory. The information about the repository usage was used to help decide whether a repository should be included in the inventory or not.

Feature	Description
Geographical Coverage	Depicts the geographical scope of the repository, depicts the coverage of the users that can upload items in the repository
ERC panel(s)	Identifies the specific ERC panel(s) covered by the repository
Content type	Presents the type of digital objects that are deposited in the repository
Access to repository	Access to the repository defining whether users can access the repository contents in general
Access to outputs upload	Access to upload research outputs to the repository
OpenAIRE Compliance	Provides information about the repository compliance or registration status in OpenAIRE

Table 5: Repository features related to repository usage

Trusted Repositories

Certification or Community endorsement

These features reflect the certification owned by the repository or its endorsement by the research community. Certification or community endorsement features are described in Table 6.

As explained in the previous sections, community endorsement has no formal definition. We considered for this study a repository endorsed by a community when this endorsement came from a disciplinary or thematic context.

Moreover, as per the trusted repository definition, we identified repositories with a community endorsement as trusted in case they were classified as domain-specific, thus did not cover all ERC-panels.

Finally, for what concerns international recognition, we considered this to be fulfilled if the repository was included in at least one of the international registries included as sources of this study.

Feature	Description
Certification	Includes information on available certification owned by the repository
Community Endorsement	Indicates whether the repository is endorsed by a specific research community
Source of Community Endorsement	Indicates the source the information about community endorsement was derived from

Table 6: Repository features related to certification or community endorsement

Essential characteristics for trusted repositories (in particular but not only for institutional or general-purpose repositories)

If no certification exists and the repository is not an internationally recognised subject/domain repository that has been endorsed by the research community, the following features will be assessed, in line with the HE AGA definition of trusted repository. The features are grouped into three categories as depicted in Table 7.

Feature	Description
1. Online policy for preservation, curation, and security of content	
Public policy for preservation, curation and security of the content	Highlights the presence of a public online policy for the repository
Policy URL	Provides a direct link to the public online policy if available
2. (Open) access and PID assignment for the content	
Metadata contains ‘licence’ field	Indicates whether the repository allows the user to set a suitable licence for the record in the metadata
PID Assignment	Indicates whether the repository allows the user to assign a persistent identifier to the contents
3. Specific metadata requirements	
Machine-actionable	Indicates if the metadata are machine-actionable
Standardised	Indicates if the repository uses standard metadata

Table 7: Repository features related to essential characteristics for trusted repositories

Alignment with the metadata requirements and recommendations in the HE MGA

In order to guide the grantees in the choice of the most suitable repository for their output, the repositories were also assessed concerning their capacity to respond to the metadata requirements and recommendations in the HE MGA.

For this purpose, the following features were identified, distinguishing between mandatory (M) and recommended (R) ones, as per the HE MGA.

Features related to metadata are described in Table 8.

We considered the Funding Stream per definition of OpenAIRE, where information regarding the Funding programme (FP7, H2020, Horizon Europe) is provided. Also, we included the information about the Funder, as repositories typically need to report funders other than the European Commission.

Feature	Description
Linked resources - (M)	Indicates if the metadata allows users to include links to or PIDs of related research outputs/tools/instruments
CC0 or equivalent Metadata - (M)	Indicates if the metadata are open under a CC0 Public Domain Dedication or an equivalent copyright waiver
Author(s) - (M)	Indicates if the metadata allow to provide information about the output Author(s)
Description - (M for data)	Indicates if the metadata allow to provide a description of the output
Title - (M)	Indicates if the metadata allow to provide a Title for the output
Date of publication / date of deposit (for research data) - (M)	Indicates if the metadata allow to provide the publication/deposition date of the output
Venue of publication/deposit - (M)	Indicates if the metadata allow to provide a venue of publication/deposit of the output
Embargo - (M for data)	Indicates if the metadata allow to provide an embargo for the output access
Funder - (M)	Indicates if the metadata allow to provide information about the Funder of the associated Grant, for example 'European Union'
Funding Stream - (M)	Indicates if the metadata allow to provide information about the funding stream for the associated Grant, for example 'Horizon Europe'
Project name - (M)	Indicates if the metadata allow to provide information about the associated Project name
Project acronym - (M)	Indicates if the metadata allow to provide information about the associated Project acronym
Grant number - (M)	Indicates if the metadata allow to provide information about the associated Grant number
Record PID - (M)	Indicates if the metadata allow to provide information about the PID of the output
Author(s) PID - (M)	Indicates if the metadata allow to provide information about the PID of the Author(s)
Organisation PID - (R)	Indicates if the metadata allow to provide information about the PID of the Authors' Organisation(s)
Grant PID - (R)	Indicates if the metadata allow to provide information about the PID of the associated Grant

Table 8: Features related to the metadata requirements and recommendations in the HE MGA.

When analysing the metadata structure of the repository, we looked for specific metadata for each item in Table 8. Some of the survey responses highlighted that often a repository has a free text field to describe a group of features in Table 8 (e.g., the granting information). However, this was not considered as structured information and cannot be considered machine actionable.

The curated results of the survey and information collection that were used as source for the study and the related analysis, as well as the inventory of identified trusted repositories are available as annexes to this report in .xlsx format. The inventory contains two spreadsheets, one for data and one for literature repositories; catch-all repositories have been included in both subsets to allow for easier consultation.

Having the repository variables visible in the inventory allows grantees to check alignment with the requirements in the HE MGA for repositories not listed in the inventory and/or allows the inventory to be further expanded/updated over time.

5. Analysis of the repositories selected for the study

This section provides an analysis of the information collected for repositories included in the study. All the 220 repositories included were tested for their readiness to facilitate compliance with the HE MGA mandates.

In the following, we show the results obtained by analysing the information collected; in some investigations we used the total of 220 repositories included in the study, in other instances we have based our discussion on the sole inventory of identified trusted repositories (165 repositories).

Repository coverage by ERC panels

To assess the coverage of repositories in terms of ERC panels, our survey asked repository managers to select every ERC panel that applied to the content that could be deposited in the repository. For completely non-responding repositories the expert group manually assessed each repository; if individual optional questions were left empty among responding repositories answers to these were not sought out by the expert group. The ERC panel structure 2021-22 (revised)[66] was used as a reference. When selecting the panels, each respondent was asked to think of the communities that can deposit outputs in the repository, and not to refer to panels that can reuse the results in the repository. All panels were selected when a repository is general-purpose (not associated with specific ERC panels). A total of 14 managers skipped this question.

Figure 5 shows the distribution of the various repositories across the different ERC panels. We highlighted in dark blue those general repositories that host outputs from all ERC panels. This figure clearly shows that some of the ERC panels have fewer available repositories than others. Not surprisingly, the panels in the Life Science domain (LS) are very well served with repositories well distributed across the different panels. The ERC panels within the Social Sciences and Humanities (SH) domain are well represented, perhaps indicating repositories with a remit combining more than one panel. The panels with the fewest repositories available are from the Physical Sciences and Engineering (PE) domain; however, this domain includes ERC Panel PE10: Earth System Science, which is one of the best panels in terms of availability of dedicated repositories. These findings confirm the information that was retrieved and described in the General overview of the repository landscape.

[66] https://erc.europa.eu/sites/default/files/document/file/ERC_Panel_structure_2021_2022.pdf

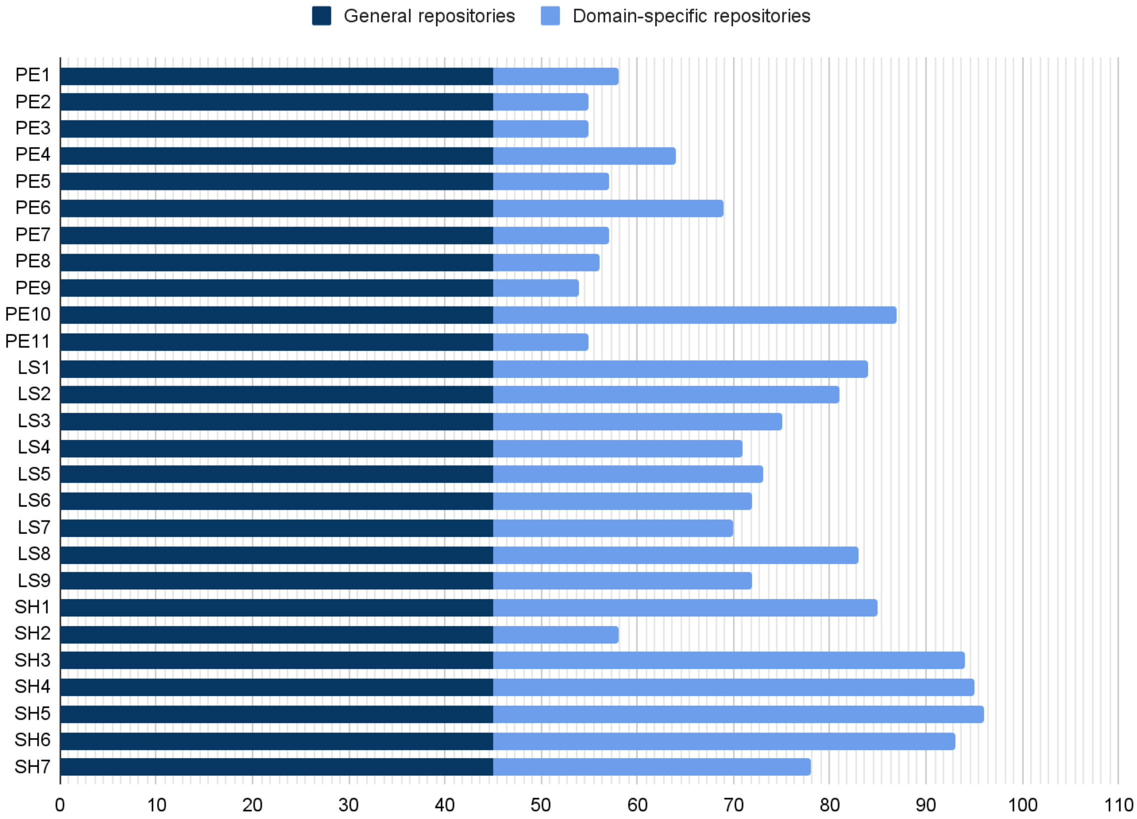


Figure 5. Coverage of ERC panels for the repositories included in the study. General (catch all) repositories are depicted in dark blue to show the existence of repositories that serve specific ERC panels (light blue).

There are a number of possible explanations for these results.

The European Strategy Forum on Research Infrastructures (ESFRI)[67] played a key role in the development of domain-specific repositories, especially those meeting the needs of ERC panels in the Life Sciences and Social Sciences and Humanities domains. Indeed, most of the repositories in the LS and SH domains in this study are linked to ELIXIR, BBMRI and CLARIN research infrastructures. It is worth noting that panel SH2: Institutions, Governance and Legal Systems does not follow the trend within the overall SH domain. This is one of the fields identified as having a gap in the ESFRI Roadmap in 2021[68].

ERC Panel PE10: Earth System Science can count on specific services from the ESFRI EPOS (the European Planet Observing System)[69] research infrastructure and ESFRI CLUSTER ENVRI-FAIR[70], which have contributed to both cultural and technical initiatives to increase repositories in this field.

[67] https://www.esfri.eu/

[68] https://roadmap2021.esfri.eu/landscape-analysis/

[69] https://www.epos-eu.org/

[70] https://envri.eu/home-envri-fair/

This panel is also one of the fields of research where a culture for data management and sharing is more widely embedded in the research workflows. Members of the Earth System community often base their work on large amounts of data (such as satellite data).

ESFRI research infrastructures do not cover all the ERC panels: there are areas where research infrastructures are less developed or missing; this is the case, for example, of the ERC panel PE1: Mathematics.

Researchers from those ERC panels that are underrepresented in this study can deposit their data and literature in general-purpose trusted repositories, such as Zenodo, or institutional trusted repositories that fulfil the requirements related to metadata.

From the ERC panel coverage, we can derive if the repository scope is general, i.e. it covers all ERC panels), or if it is domain (or thematic) specific. These latter repositories cover more than one but not all of the ERC panels. Only 20% of the repositories in this study are general-purpose. The results of this exercise are reported in Table 9.

	Number of repositories	Percentage
General repositories	45	20%
Domain-specific repository	161	73%
NA	14	7%

Table 9. Scope of the repositories in terms of specification of their contents: general (purpose) repositories and domain-specific (or thematic) repositories; this information is not available for 14 repositories in the study.

Geographical scope of the repositories

Geographical scope of repository	Number of repositories	Percentage
International	122	55%
National	33	15%
Institutional	59	27%
N/A	6	3%

Table 10. Geographical scope of the repositories included in the study. The table depicts the coverage (International, National or Institutional) with respect to the capacity to upload digital objects in the repository; the table shows number of repositories falling in each category, as well as corresponding percentage.

The geographical scope of a repository depicts the location of its depositors. A repository may accept uploads from anyone (international), from individuals with certain institutional affiliations (institutional) or connections to organisations based in a specific country (national). Table 10 shows the classification of repositories in the study in terms of geographical scope. It was not possible to determine the geographical scope for six repositories, mainly due to the lack of response from repository managers to the corresponding question in the survey.

It is interesting to note that the majority of repositories in the study have an international scope (55%) and, as depicted in Table 11, two thirds of these declare to be or are identified as endorsed by a specific thematic or disciplinary community (66%). 53% of these international repositories host research data.

These findings reflect the need for most thematic and disciplinary communities to engage in international collaboration for their work. These repositories tend to host data irrespective of the geographical origin of the researchers or their affiliation.

The results also highlight that the management and preservation of research data can best be sustained if a repository serves a larger international community, rather than an institutional or national audience.

Focus: repositories with International coverage		
	Number of repositories	**Percentage**
Content type		
Data	65	53%
Literature	29	24%
Catch all	28	23%
Endorsement by thematic/disciplinary research community		
Yes	81	66%
No or N/A	41	34%

Table 11. Geographical scope of the repositories included in the study. The table reports information about the content type and community endorsement for repositories with an international coverage.

Repository contents

The repositories included in the study were classified on the basis of the type of content they host. To this aim, the different content types were grouped into four categories:

- Literature - articles, books, conference proceedings, and text formats in general
- Data
- Software/Code
- Other - media, lectures, etc.

Type of content accepted	**Number of repositories**	**Percentage**
Literature - articles, books, conference proceedings, etc...	120	55%
Data	166	75%
Software/Code	70	32%
Other - media, lectures, etc...	89	40%

Table 12: Repository contents. The table reports the different types of contents that repositories in the study can host. As multiple choices could be provided, the numbers do not sum up to the total number of repositories.

The results of this mapping exercise are reported in Table 12. The percentages presented in the table reflect the nature of the candidate repositories for the study, where 60% of the repositories were derived from data repository registries, while 40% came from registries focusing on literature repositories.

It is worth noting that the categories ‘Software/Code’ and ‘Other’ are present in one third of the repositories in the study, even if they never appear as the sole content, as they are always associated with either literature, data or both.

Given the responses, the repositories were divided into three main categories:

- Catch All - repositories that can host both data and literature, and that are sometimes suitable also for software and other outputs
- Data (no Literature) - repositories host data but not literature, and are sometimes suitable also for software and other outputs
- Literature (no Data) - repositories host literature but not data, and are sometimes suitable also for software and other outputs.

The results of this grouping exercise are detailed in Table 13.

Repository type by content	Number of repositories	Percentage
Catch All - Data & Literature	66	30%
Data (no Literature)	100	45%
Literature (no Data)	54	25%

Table 13. Repositories categorisation in terms of content type. The table reports repositories categorised as Catch all, Data (no Literature) and Literature (no Data).

Repository metadata availability

Here, we summarise to what degree the repositories in the study are capable of providing metadata for uploaded objects. Figure 6 below displays the full overview of all repositories in the study and demonstrates the broad support for many of the various metadata attributes. Almost all repositories contain metadata for authors, title, date of publication/deposit, and a description of the object. With regards to PIDs, the record PID (e.g., DOI/Handle) is the most widely supported with 80%, followed by author PIDs (e.g., ORCID) with 60%. These findings also indicate that 73% of the repositories studied provide their metadata with a CC0 or similar public domain dedication.

It is interesting to highlight that only 38% of the repositories in the study provide structured metadata related to the funding stream and only 38% provide structured information on the project acronym.

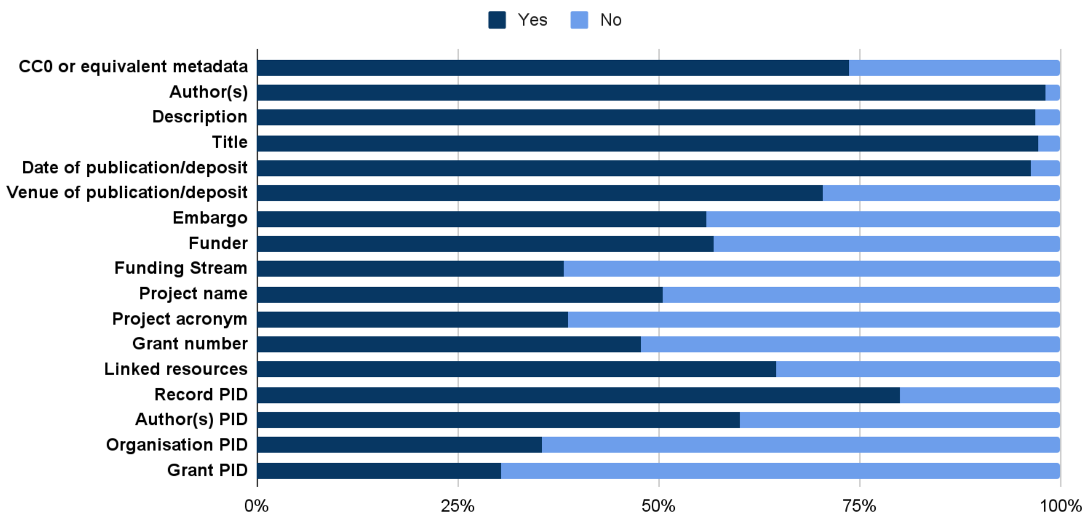


Figure 6: Metadata availability for all 220 repositories in the inventory

If we conduct the same analysis but focus on only those repositories categorised as trusted, i.e., the 165 repositories in total, the overall picture does not change, as visualised in Figure 7.

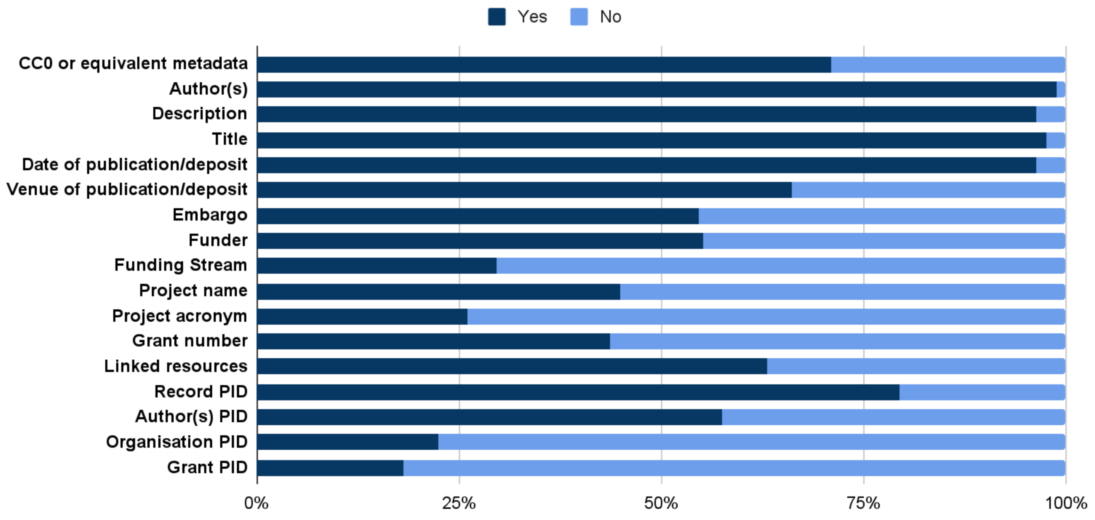


Figure 7: Metadata availability for only the 165 trusted repositories in the inventory

It is worth noting that metadata fields for Organisation PIDs and Grant PIDs are least present in the repositories studied, whereas fields for Author PIDs are present in more than half of all cases. Not surprising, metadata fields for Record PIDs are present in the vast majority of repositories analysed.

In addition, we note that the fields related to grant information are mainly free text and thus lack machine-actionable, interoperable and standard metadata.

Openness to deposition

Approximately 70% of the repositories studied allow external users to upload resources in either an open or restricted way. The remaining 29% that are closed to deposition are mostly institutional, while those with restricted deposition are mostly domain-specific repositories that curate uploaded content or handle content delivery through more manual processes, where researchers need to get in touch with a repository manager to arrange for deposition. Table 14 reports repository distribution in terms of openness to deposition.

Access to data upload	Number of repositories	Percentage
Open: external users can upload the content to the repositories with no barriers.	44	20,0%
Restricted: external users can overcome access barriers (eg: the repository requires authentication/authorisation to upload content)	111	50,5%
Closed: external users cannot overcome access barriers (eg: only internal users can upload content to the repository)	64	29,1%
N/A	1	0,5%

Table 14. Repositories' openness to deposition. This table uses the re3data categorisation.

Openness of the contents

As depicted in Table 15, most of the repositories studied provide open access to their content, whereas 13% require some kind of authentication or authorisation. Some repositories only grant data access after individual application, while some repositories contain mixed materials where some are open and others may be limited only to institutional members.

Access to repository	Number of repositories	Percentage
Open: external users can access the content of the repositories with no barriers	191	87%
Restricted: external users can overcome access barriers (e.g.: the repository requires authentication/authorisation to access the content)	29	13%

Table 15. Repositories' openness of the contents. This table uses the re3data categorisation.

Curation, preservation, and security policies

154 of the 220 repositories were found to have some kind of policy for preservation, curation and security of their content online, and many others reported that this was work in progress and would be coming in the near future. Unfortunately, even when a policy was present, there was no standardisation in policy structure or presentation, nor in their focus, process, and commitments. This is something we hope repository managers will take note of.

Repositories that publish their policies tend to do so within sub-sections of their websites or on the website of the organisation managing the service. To compound matters, these sections can be labelled with different names such as 'policy', 'terms and conditions', 'F.A.Q.', and it is interesting to note that the information about the key aspects of the policy was often fragmented in different sections. This situation does not facilitate the easy assessment of policy and, as a consequence, the detection of trusted repositories.

The percentage of repositories that have an online policy varies when we distinguish them by geographical coverage, with International repositories more likely to have a policy available online. Table 16 reports on the existence of a public policy depending on the repositories' geographical scope.

	Repository with public policy	Percentage
International	92	75%
National	25	76%
Institutional	32	54%
N/A	5	83%

Table 16: Repositories that offer a publicly available policy on curation, preservation and security of contents, differentiated by geographical scope. Percentages are given considering the total number of repositories in each category and not with respect to the total number of repositories in the study.

Table 17 reports on the availability of repositories with public curation, preservation and security policies among the general and domain-specific repositories respectively. 72% of domain-specific repositories have a public policy online, while this is the case for 69% of general repositories.

	Repository with public policy	Percentage
General repositories	31	69%
Domain-specific repository	116	72%
N/A	7	50%

Table 17. Repositories that offer a publicly available policy on curation, preservation and security of contents, differentiated by repository type (general or domain-specific repositories). Percentages are given considering the total number of repositories in each category and not with respect to the total number of repositories in the study.

OpenAIRE compliance

OpenAIRE compliance for the repositories included in the study was derived from the OpenAIRE website and is provided in Table 18.

While a large number (65) of literature repositories are compatible with the OpenAIRE guidelines, only 9 data repositories were considered compatible; Table 19 shows the percentage of repositories not yet registered in OpenAIRE, differentiated by content type. Percentages are given considering the total number of repositories in each category. It is noteworthy that 45% of the repositories studied are not yet registered in OpenAIRE.

Type of OpenAIRE compliance	Number of repositories
OpenAIRE 3.0 (OA, funding)	18
OpenAIRE Basic (DRIVER OA)	29
OpenAIRE 2.0+ (DRIVER OA, EC funding)	14
OpenAIRE PubRepos v4.0	4
OpenAIRE Data	9
Collected from a compatible aggregator	34
Not yet registered	98
Information not available	14

Table 18: OpenAIRE compliance of repositories in the study.

Number of repositories not yet registered in OpenAIRE by Content type	Number of repositories	Percentage
Data (no literature)	63	63%
Literature (no data)	9	17%
Catch All (Data and Literature)	26	39%

Table 19: Repositories not yet registered in OpenAIRE, differentiated by content type. Percentages are given considering the total number of repositories in each category and not with respect to the total number of repositories in the study.

Trusted repositories overview

	Certified repository	Community Endorsement	Essential Characteristics	Trusted repository count
General repositories	11	0	21	24
Domain-specific repository	82	104	75	130
N/A	10	4	6	11
Total	103	108	102	165

Table 20: Trusted repositories overview. The table reports the number of repositories that were identified to adhere to the definition of trusted repository given in the HE AGA, with a focus on the type of repository (domain-specific or general-purpose). The table also reports the reason why the repository was identified as trusted: certification, community endorsement for domain/thematic repository and/or essential characteristics as explained in the Concepts section.

Based on the definition in the HE AGA, repositories can be classified as trusted for a number of reasons, as the certification, community endorsement and essential characteristics depicted in the Concepts section may all apply to a repository. Of the 220 repositories included in the study, 165 can be considered trusted. The first reason to consider a repository as trusted is community endorsement. 108 /165 repositories were reported to be endorsed by a community.

As reported in Table 20, the vast majority of domain-specific repositories that can be considered trusted are considered so due to endorsement from their community. Given the HE AGA definition, no general repositories can be considered to have community endorsement and therefore be trusted. However, a number of general repositories were found to fulfil the essential characteristics depicted in the HE AGA, and therefore were considered trusted.

Of the 165 trusted repositories, 62% are certified repositories, mostly due to CTS certification (83 repositories in the study were endorsed with CTS at least once in their lifespan). The high percentage of repositories with a CTS certification is reflective of the strategy used for repository inclusion in the study (see the Methodology section).

It is also worth noting that the CTS certification was developed in the Open Science context. It is therefore expected that most of the certified repositories fall under the CTS (6 general and 71 domain-specific, and 6 repositories for which we could not determine the ERC panel coverage).

As depicted in Table 21, only 20 institutional repositories included in the study are certified, with adherence to essential characteristics being the first reason for considering an institutional repository trusted (21 repositories). 80% of the trusted international repositories in the study have been endorsed by a research community.

	Certified repository	Community Endorsement	Essential Characteristics	Trusted repository count
International	54	80	56	100
National	27	19	22	29
Institutional	20	9	21	32
N/A	2	0	3	4
Total	103	108	102	165

Table 21: Trusted repository overview with respect to geographical scope of the repository. The table reports the number of repositories that were identified to adhere to the definition of trusted repository given in the HE AGA. The table also reports the reason why the repository was identified as trusted: Certification, Community endorsement for domain/thematic repository and essential characteristics as explained in the Concepts section.

Some of the repositories fulfil more than one criterion for being trusted. This is well represented in Figure 8, where a Venn diagram details the reasons for repositories to be considered trusted. Pertinently, 39 certified repositories out of 103 do not meet essential characteristics as described in the HE AGA. As the Horizon Europe AGA requirements were published quite recently, these findings illustrate that certification criteria do not fully meet the HE AGA requirements and therefore should not be considered a proxy for trusted repositories as identified in the HE AGA. It may take some time to embed these new requirements into certification standards. The number of repositories that can be considered trusted due only to community endorsement is higher than the number of repositories that only meet certification or essential characteristics criteria. This result may suggest that the criteria related to community endorsement should be updated or enhanced in the future.

Overall, only 21 repositories out of 165 can be considered trusted due to both certification and essential characteristics fulfilment without being community endorsed, 23 meet both certification and community endorsement criteria without fulfilling the essential criteria, and 18 are trusted both because they fulfil essential characteristics and they are endorsed by a research community without being certified. 43 repositories were identified to meet all three criteria for trusted repositories.

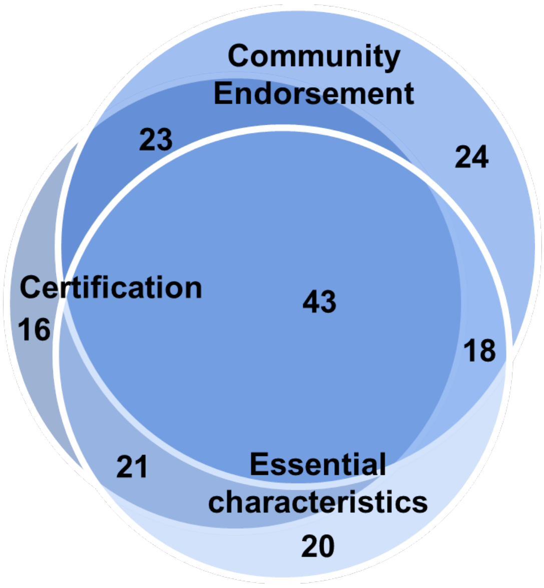


Figure 8 Venn Diagram of trusted repositories depicting the reasons for being considered trusted. The Venn diagram shows clearly how some repositories fulfil more than one criterion (certification, community endorsement, and essential characteristics).

Of the 220 repositories selected to be included in the study, 55 did not meet any of the criteria to be considered trusted.

Apart from the certification and community endorsement, which are Yes/No criteria, we analysed the most frequent reasons for not meeting the essential characteristics for trusted repositories. For those 55 repositories that were neither certified nor endorsed and whose acceptance as trusted repositories thus depended on meeting the essential characteristics, results are shown in Table 22. The lack of a public policy for preservation, curation and security of the contents is the most frequent reason for not displaying all the essential characteristics of trusted repositories.

	Policy missing	Licence field missing in metadata	No PID assignment	No Machine Actionable Metadata	No Standard Metadata
Number of repositories	35	14	9	8	8

Table 22: Reasons for not being trusted by failing to meet the essential characteristics for trusted repositories. Numbers do not sum up to the total number of non-trusted repositories because some repositories fail to meet more than one requirement. Analysis for repositories that are neither certified nor endorsed.

Of the trusted repositories that were either certified or endorsed, some failed to comply with one or more essential characteristics for trusted repositories. In particular, as reported in Figure 8, 39 certified repositories and 47 community endorsed repositories did not meet the essential characteristics, we analysed the reason why these repositories failed to meet them and report the results in Table 23.

	Policy missing	Licence field missing in metadata	No PID assignment	No Machine Actionable Metadata	No Standard Metadata
Certified repositories missing essential characteristics	21	15	5	7	6
Community endorsed repositories missing essential characteristics	21	24	4	6	9

Table 24: Reasons for trusted repositories (either certified or endorsed by the community) failing to meet the essential characteristics for trusted repositories. Some repositories fail to meet more than one requirement.

Results in Table 24 basically confirm that the main reason not to meet the essential characteristics for trusted repositories is the lack of clear policy and metadata licence.

In the Conclusions, we will discuss some of the motivations for this finding.

Repository readiness to facilitate compliance with the Horizon Europe MGA metadata requirements

We analysed in detail the readiness of the repositories included in this study to facilitate compliance with the Horizon Europe MGA metadata requirements. The vast majority (over 90%) of those repositories meet the basic metadata requirements, allowing research outputs to be described with at least basic information (Author(s), Description, Title, Date of publication/deposit), as described in the Concepts section.

Repository readiness has been categorised by looking at the number and type of metadata that are needed by the grantees to comply with the requirements in the HE MGA. Following Table 3, repositories can therefore be classified as ready to comply with Mandatory and Recommended, with only Mandatory and with only Basic requirements. In some cases, information collected in this study was not enough to define the readiness level of the repository.

We would like to stress that when a trusted repository is not in line with all the mandatory metadata requirements, other services may support ERC grantees in providing all the necessary information to the funder, even if this is not done via the repository directly. This of course does not modify the readiness of the repository in terms of facilitating the compliance with the HE MGA, but still enables appropriate reporting. For example, OpenAIRE offers a ‘Link’ function through which it is possible to directly provide information about the funding for a given output that is already deposited in one of the repositories registered in OpenAIRE. Information provided is sent directly to the EC participant portal grant management service to be shown in the project continuous reporting section. The OpenAIRE ‘Link’ service also allows authors to provide information about linked resources and research outputs that are needed to validate the results uploaded in a repository.

Table 24 shows the readiness level of the repositories in the study to meet HE MGA metadata requirements. Only 2% of repositories in the study fulfil all mandatory metadata requirements, whereas none meet all recommended and mandatory metadata requirements.

Repositories selected for the study	Mandatory and recommended	Mandatory	Basic	Data inconclusive
Number of repositories	0	4	200	20

Table 24: Readiness of repositories in the study to facilitate compliance with the Horizon Europe MGA metadata requirements. Note that the numbers don’t add up to the total number of repositories in the study as the repositories that fulfil the mandatory requirements also fulfil the basic ones.

We also assessed the level of readiness for trusted repositories and reported results in Table 25, where we can see a similar distribution.

Repositories identified as trusted	Mandatory and recommended	Mandatory	Basic	Data inconclusive
Number of repositories	0	3	151	14

Table 25: Readiness of trusted repositories in the study to facilitate compliance with the Horizon Europe MGA metadata requirements. The numbers don’t add up to the total number of repositories in the study as the repositories that fulfil the mandatory requirements also fulfil the basic ones.

The four repositories that fulfil all mandatory metadata requirements are:

- Zenodo,
- Hyper Article en Ligne,
- OAPEN, and
- <intR>²Dok.

Not surprisingly, the first three repositories enjoy a very close relationship with Funders. In the following, we briefly analyse the four repositories as well as those that are recommended by the ERC Scientific Council in the ‘Open Access Guidelines for research results funded by the ERC’ document. Namely, Europe PMC for publications in the Life Sciences and arXiv for those in relevant areas of the Physical Sciences and Engineering.

Name	Zenodo
Description and scope	Zenodo is the open access repository created in collaboration between OpenAIRE and CERN and was born with the aim to support the European Commission and its Open Science policies. It is a general-purpose, international catch all repository and can thus be used to deposit both data and literature as well as software through a collaboration with GitHub.
Trusted repository	Zenodo has been identified as trusted because it fulfils all the essential characteristics required - policy, (open) access and PID assignment, metadata requirements. Zenodo is not certified.
Readiness to facilitate the compliance with HE MGA	Zenodo meets all the mandatory metadata requirements. It does not support PIDs for Organisations and associated Grants. Therefore, it does not allow grantees to provide the additional recommended metadata specified in the HE MGA. However, Zenodo provides a specific set of metadata to describe and provide PIDs for linked resources to validate the results of the items deposited.

Name	Hyper Article en Ligne (HAL)
Description and scope	Hyper Article en Ligne (HAL) is the open access repository developed and managed by CCSD (Centre pour la Communication Scientifique Directe), a support and research unit (UAR 3668) of the CNRS. HAL is the French national open archive for literature. It is a general-purpose repository and can also provide specific services for code preservation thanks to a collaboration with Software Heritage. It offers more than 130 institutional portals from universities and other research organisations in France. HAL is included in the Second French National Plan for Open Science to be further developed.
Trusted repository	HAL has not been identified as a trusted repository. HAL did not meet the essential criteria for a trusted repository because it does not provide a public policy for preservation, curation and security of its contents.
Readiness to facilitate the compliance with HE MGA	HAL meets all the mandatory metadata requirements. It does not support PIDs for Organisations and associated Grants. Therefore, it does not allow grantees to provide the additional recommended metadata specified in the HE MGA. HAL does not provide a specific metadata field to describe and provide PIDs for linked resources to validate the results of the items deposited.

Name	OAPEN Library
Description and scope	OAPEN (Open Access Publishing in European Networks) is managed by the OAPEN Foundation, a not-for-profit organisation based in the Netherlands, with its registered office at the National Library in The Hague. OAPEN is dedicated to open access peer-reviewed books and covers all ERC Panels. Like Zenodo, OAPEN was developed thanks to a project co-funded by the EU. OAPEN is recommended by the ERC Scientific Council in its ‘Open Access Guidelines for research results funded by the ERC’ for book chapters as well as long-text publications such as monographs or edited collections.
Trusted repository	OAPEN has been identified as trusted because it fulfils all the essential characteristics required - policy, (open) access and PID assignment, metadata requirements. OAPEN is not certified.
Readiness to facilitate the compliance with HE MGA	OAPEN meets all the mandatory metadata requirements. It does not support PIDs for Organisations and associated Grants. Therefore, it does not allow grantees to provide the additional recommended metadata specified in the HE MGA. OAPEN does not provide a specific metadata field to describe and provide PIDs for linked resources to validate the results of the items deposited.

Name	<intR>²Dok
Description and scope	The disciplinary Open Access repository <intR>²Dok (pronounced: ’Inter-Zwei-Dok’) is the central publication platform of the specialist information service for international and interdisciplinary legal research set up by the German Research Foundation at the Berlin State Library - Prussian Cultural Heritage. It is a national repository specialised in ERC Panels SH2: Institutions, Governance and Legal Systems and SH3: The Social World and Its Diversity.
Trusted repository	<intR>²Dok has been identified as trusted because it is certified (DINI certificate) and it fulfils all the essential characteristics required - policy, (open) access and PID assignment, metadata requirements. Despite its thematic dimension, it was not possible to assess if the repository is endorsed by a relevant research community.
Readiness to facilitate the compliance with HE MGA	<intR>²Dok meets all the mandatory metadata requirements. It does not support PIDs for associated Grants and therefore does not allow grantees to provide the additional recommended metadata specified in the HE MGA. <intR>²Dok provides a specific metadata field for linked resources to validate the results of the items deposited.

Name	ArXiv
Description and scope	arXiv (pronounced ’archive’, the X represents the Greek letter chi χ) is a free distribution service and an open access archive for scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. It is developed and managed by Cornell University. arXiv was the first and is one of the most well-known preprint servers. Elsevier allows authors of articles published in their journals to update preprints that have already been posted on arXiv with the postprint (author’s accepted manuscript), without embargo period.
Trusted repository	arXiv has been identified as trusted because it fulfils all the essential characteristics required - policy, (open) access and PID assignment, metadata requirements - and it is commonly used and endorsed by the research communities, and internationally recognised. arXiv is not certified.
Readiness to facilitate the compliance with HE MGA	arXiv does not meet most mandatory metadata requirements specified in the HE MGA, and it does not provide the additional recommended metadata either. arXiv does not meet the basic metadata requirements identified in this study. It is not possible to provide information related to funding via the repository nor to provide PIDs, except for the one of the upload. arXiv does not provide a specific metadata field to describe and provide PIDs for linked resources to validate the results of the items deposited.

Name	Europe PMC
Description and scope	Europe PMC provides comprehensive access to Life Sciences literature from trusted sources. Europe PMC is a literature repository that hosts publications, preprints and other documents enriched with links to supporting data, reviews, protocols, and other relevant resources. It is hosted by EMBL’s European Bioinformatics Institute (EMBL-EBI) and part of the ELIXIR infrastructure. Europe PMC is partnered with PubMed Central (PMC), and endorsed and supported by a group of international science funders, including ERC, as their repository of choice.
Trusted repository	Europe PMC has been identified as trusted because it is commonly used and endorsed by the research communities, and internationally recognised. Europe PMC is not certified. It was not possible to assess if a public policy is available for preservations, curation and security of its contents.
Readiness to facilitate the compliance with HE MGA	Europe PMC meets most but not all mandatory metadata requirements specified in the HE MGA (it is not possible to provide the grant acronym). It does meet the basic metadata requirements identified in this study. It also provides the additional recommended metadata. Europe PMC metadata are not open access under a Creative Commons Public Domain Dedication (CC0) or equivalent.

Conclusions

Overall, the repository inventory and the analysis provided in the previous sections demonstrate that the research outputs within all the ERC panels are catered for, both through general and domain-specific repositories. In total, we identified 165 repositories that fulfil the HE MGA definition of a trusted repository. The vast majority (more than 90%) of the repositories in this study are found to be in line with the basic requirements in terms of metadata. However, there is a surprisingly low number of repositories that have enabled the registration of all the metadata ascribed to the HE MGA 'Mandatory’ (2%) category. Even some of the most well-established and popular repositories miss out on one or several of the included items.

Upon close inspection of the different repositories included in this study, one persistent theme was their diversity, particularly for those domain-specific data repositories where the needs of different sub-disciplines have strongly shaped the technical solutions implemented. Changing and implementing extended metadata fields and standards over time can be very difficult if the repository software is not generic or off-the-shelf, but rather built as a bespoke solution to serve the needs of a very specific community. Thus, the tension between domain-agnostic standardised funder requirements and the heterogeneous systems that the research community has built is likely to remain for the time being. As mentioned in the previous paragraph, most repositories included in the study are not sufficiently ready to enable grantees to be compliant with the HE MGA open science requirements.

It is common for research institutions to have their own institutional repositories, something which the basic growth statistics from the earlier parts of the report have already shed light on. In this study we could only include a small share of such repositories as their number is so high. Of the 59 institutional repositories included in the study, 32 fulfilled the HE MGA requirements for a trusted repository.

This report and the associated inventory lay the foundations for ERC grantees to be able to more easily compare, assess and choose an appropriate repository for data or literature deposition. The inventory should be considered an inexhaustive snapshot of the landscape. Like any digital tool, the inventory and the resultant list of trusted and nearly-trusted repositories are only as good as the methodology used to define them. Yet, we believe it to be a sufficient baseline of repositories across the different ERC panels. The inventory and associated overview also provide an indication as to where potential future development efforts in coverage should be directed.

The repository landscape is evolving rapidly, as are funder requirements for data and repository metadata. We believe repositories would benefit from making more comprehensive information pages about their policies, licences, and metadata standards available, in both a human and machine-readable fashion. The research community as a whole may benefit too, were we able to come together and agree on a standardised notation for these metadata across repositories.

We believe our experience in this project bears resemblance to the experience that grantees may have when trying to interpret funder requirements, given the complex reality of the repository landscape. Many concepts such as community endorsement or what counts as a sufficient preservation policy are hard to gauge even for individuals conducting a dedicated investigation, let alone for individual grantees that do not have the same amount of time and effort to spare on exploring the matter. The differing requirements for different types of repositories (general versus discipline-specific) is also something that is not always a perfect fit with the diverse reality of repositories, as there can be, for example, important national disciplinary archives that could be appropriate for use but that may not be what can be interpreted as ‘internationally recognised’.

Currently, some of the most crucial components of trust - the preservation, curation and security of repository contents - are very hard to assess in a transparent and fair manner from the outside. There are as yet no widely adopted, externally verifiable, policies or mechanisms that ensure all repository contents are maintained and preserved to the highest standards. Here, we would urge stakeholders in the repository space to advance the development and adoption of standardised policies and practices so that trust can be built on a solid foundation.

We consider this work to be the starting point for stakeholders and other interested parties to take further, to update, refine and improve over time. We hope that it may be used for discussions on metadata harmonisation between repositories, for gap analyses on the repository landscape, and to promote standards for FAIR and open research data and literature repositories.

References

Cannon, M. et al. (2021). Repository Features to Help Researchers: An invitation to a dialogue. Zenodo. https://doi.org/10.5281/zenodo.4683794

Corrado, E. M. (2019). Repositories, Trust, and the CoreTrustSeal. *Technical Services Quarterly*, 36(1), 61–72. https://doi.org/10.1080/07317131.2018.1532055

Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., de Vries, J., L'Hours, H., Davidson, J., & Angus W. (2020). FAIRsFAIR Data Object Assessment Metrics (0.5). Zenodo. https://doi.org/10.5281/zenodo.6461229

Drysdale, R., McEntyre, J., Durinx, C., and Blomberg, N. (2018). The Process for the Selection of ELIXIR Core Data Resources [version 1; not peer reviewed]. *F1000Research* 2018, 7(ELIXIR):1712 (document) https://doi.org/10.7490/f1000research.1116248.1

European Commission (2017). H2020 Programme Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020 - Version 3.2, 21 March 2017. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

European Commission (2021). Horizon Europe (HORIZON) - Annotated Model Grant Agreement. Version 0.2, 30 November 2021. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/aga_en.pdf

European Commission (2022a). Horizon Europe (HORIZON) - Model Grant Agreement. Version 1.1., 15 April 2022. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/agr-contr/general-mga_horizon-euratom_en.pdf

European Commission (2022b). European Research Data Landscape: final report, Publications Office of the European Union. https://data.europa.eu/doi/10.2777/3648

European Research Council Scientific Council (2022). Open Research Data and Data Management Plans, Information for ERC grantees, Version 4.1, 20 April 2022. https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf

Eve, M. P. (2014). *Open Access and the Humanities*. Cambridge University Press. https://doi.org/10.1017/cbo9781316161012

Ferguson, C., Araújo, D., Faulk, L., Gou, Y., Hamelers, A., Huang, Z., Ide-Smith, M., Levchenko, M., Marinos, N., Nambiar, R., Nassar, M., Parkin, M., Pi, X., Rahman, F., Rogers, F., Roochun, Y., Saha, S., Selim, M., Shafique, Z., … McEntyre, J. (2020). Europe PMC in 2020. *Nucleic Acids Research*, 49(D1), D1507–D1514. https://doi.org/10.1093/nar/gkaa994

Fraser, N., Brierley, L., Dey, G., Polka, J. K., Pálfy, M., Nanni, F., & Coates, J. A. (2021). The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLOS Biology*, 19(4), e3000959. https://doi.org/10.1371/journal.pbio.3000959

Gentil-Beccot, A., Mele, S., & Brooks, T. C. (2009). Citing and reading behaviours in high-energy physics. *Scientometrics*, 84(2), 345–355. https://doi.org/10.1007/s11192-009-0111-1

Ginsparg, P. (1994). First Steps Towards Electronic Research Communication. *Computers in Physics*, 8(4), 390. https://doi.org/10.1063/1.4823313

Hoy, Matthew B. (2020). Rise of the Rxivs: How Preprint Servers are Changing the Publishing Process, *Medical Reference Services Quarterly*, 39:1, 84-89. https://doi.org/10.1080/02763869.2020.1704597

Kindling, M., Pampel, H., van de Sandt, S., Rücknagel, J., Vierkant, P., Kloska, G., Witt, M., Schirmbacher, P., Bertelmann, R., & Scholze, F. (2017). The Landscape of Research Data Repositories in 2015: A Re3Data Analysis. *D-Lib Magazine*, 23(3/4). https://doi.org/10.1045/march2017-kindling

Larivière, V., Sugimoto, C. R., Macaluso, B., Milojević, S., Cronin, B., & Thelwall, M. (2014). arXiv E-prints and the journal of record: An analysis of roles and relationships. *Journal of the Association for Information Science and Technology*, 65(6), 1157–1169. https://doi.org/10.1002/asi.23044

Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., ... & Westbrook, J. (2020). The TRUST Principles for digital repositories. *Scientific Data*, *7*(1), 1-5. https://doi.org/10.1038/s41597-020-0486-7

Manghi, P., Atzori, C., Bardi, A., Baglioni, M., Schirrwagen, J., Dimitropoulos, H., La Bruzzo, S., Foufoulas, I., Mannocci, A., Horst, M., Czerniak, A., Kiatropoulou, K., Kokogiannaki, A., De Bonis, M., Artini, M., Ottonello, E., Lempesis, A., Ioannidis, A., Manola, N., & Principe, P. (2021). OpenAIRE Research Graph Dump (4.0). Zenodo. https://doi.org/10.5281/zenodo.5801283

Mannocci, A., Baglioni, M., & Manghi, P. (2022). “Knock knock! Who’s there?“ A study on scholarly repositories’ availability. *26th International Conference on Theory and Practice of Digital Libraries*, Padua, Italy. arXiv. https://doi.org/10.48550/arXiv.2207.12879

Manola, N., Papageorgiou, H., Grypari, I., & Lempesis, A. (2021a), *Monitoring the open access policy of Horizon 2020: final report.* Publications Office, 2021, https://data.europa.eu/doi/10.2777/268348

Manola, N., Papageorgiou, H., Grypari, I., & Lempesis, A. (2021b). MOAP Horizon 2020 database [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4899767

Mugabushaka, A.-M., Baglioni, M., Bardi, A., & Manghi, P. (2021). Scholarly outputs of EU Research Funding Programs: Understanding differences between datasets of publications reported by grant holders and OpenAIRE Research Graph in H2020. arXiv. https://doi.org/10.48550/arXiv.2109.10638

Öchsner, A. (2013). *Introduction to Scientific Publishing - Backgrounds, Concepts, Strategies.* SpringerBriefs in Applied Sciences and Technology. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-38646-6

Office of The Director, National Institutes of Health (2020). Supplemental Information to the NIH Policy for Data Management and Sharing: Selecting a Repository for Data Resulting from NIH-Supported Research. Notice Number: NOT-OD-21-016. Release Date: October 29, 2020. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-016.html

Pinfield, S., Salter, J., Bath, P. A., Hubbard, B., Millington, P., Anders, J. H. S., & Hussain, A. (2014). Open-access repositories worldwide, 2005-2012: Past growth, current characteristics, and future possibilities. *Journal of the Association for Information Science and Technology,* 65(12), 2404–2421. https://doi.org/10.1002/asi.23131

Van de Sompel, H. & Lagoze, C. (2000). The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine*, 6(2). https://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html

Vines, T. H., Albert, A. Y., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., ... & Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current biology*, 24(1), 94-97. https://doi.org/10.1016/j.cub.2013.11.014

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 160018. https://doi.org/10.1038/sdata.2016.18