How to find a trustworthy repository for your data

WHAT IS REQUIRED?

By default Horizon 2020 projects participate in the Open Research Data Pilot and they must deposit the following data in a research data repository:

1. all data needed to validate the results presented in scientific publications, including the metadata that describe the research data deposited. This is called the "underlying data". These data must be deposited as soon as possible.
2. any other data (for instance curated data not directly attributable to a publication, or raw data), including the associated metadata, as specified and within the deadlines laid down in the Data Management Plan – that is, according to the individual judgement by each project/grantee.
3. projects should also provide information via the chosen repository about the tools that are needed to validate the results, e.g. specialised software or software code, algorithms and analysis protocols. Where possible, they should provide these instruments themselves, or alternatively, provide direct access to them.

The legal requirements for participating projects are set out in article 29.3 of the Model Grant Agreement, included by default in the Grant Agreement. Details on exceptions and on - partially - opting out can be found in the Participant Portal H2020 Online Manual.

THE EC PREFERS CERTIFIED REPOSITORIES

In the Guidelines on FAIR Data Management in Horizon 2020, the European Commission states: "Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible."

Researchers, information managers and other stakeholders can rely on a framework of various international certification standards for digital repositories in order to assess and improve the quality of their work processes and management systems. "Trustworthy Digital Repository" (TDR) is a term often used in this respect.

In the European Framework for audit and certification of digital repositories three certification instruments, with increasing degrees of complexity and depth, are available:

- CoreTrustSeal (CTS): this is based on Data Seal of Approval (DSA) and World Data System (WDS). All digital repositories that have one or more of these certifications are listed at https://www.coretrustseal.org/why-certification/certified-repositories/
- Nestor Seal: verification according to DIN 31644

- ISO 16363 certification

The assessments vary in intensity from a peer review of completed documentation (self-assessment) to a prepared on-site visit by an external audit team. These instruments are used worldwide. Data sponsors, producers and re-users may trust any managing body that has been certified according to one of the above standards.

There are also data repositories with a long standing and solid user base, like Zenodo, that have no certification; you'll find other examples in OpenDOAR, the directory of Open Access repositories. It is expected that these repositories will apply for certification in the near future, because the organisation of research funding and research performing organisations in Europe (Science Europe) has developed criteria for the selection of trustworthy repositories. These criteria contain a recommendation that repositories that are not certified yet seek certification by such a body. We encourage you to support this in your conversation with repositories, simply by alerting them to this development. For the time being, there don't seem to be consequences for repositories that don't have a certification or for using such repositories.

Read the "Background" section below when you want to know how the requirements for certification support the FAIR data principles.

HOW TO COMPLY WITH THE OPEN DATA PILOT REQUIREMENTS?

During the project, you are expected to manage your data, metadata and documentation in accordance with the standards and good practices in your discipline - see the OpenAIRE RDM handbook. Before the deadlines mentioned above, you deposit this package in a trustworthy repository. Remember to also deposit data that cannot be made open. Once the data are published you refer to them with a so-called persistent identifier or PID, like a DOI or a URN:NBN. This enables the research funder and potential users of your data to find and later cite your data. Watch this video by Research Data Netherlands when you want to know more about PIDs and data citation.

Attaching a usage licence to your data tells others what they can or cannot do with the data. OpenAIRE recommends to use the Creative Commons CC0 waiver or the CC-BY licence for open access to data. Creative Commons offers more restrictive licences for data if needed, and the EUDAT License Selector helps you to select software licences as well.

The good news is that trustworthy repositories can help you with PIDs and licences.

WHICH REPOSITORY TO USE?

The general steps for finding a data repository are:

1. use a disciplinary repository if there is one;
2. alternatively, use the institutional repository, if you have one where the data will also be available for the long term;
3. use the catch-all repository Zenodo, maintained by CERN;
4. or search the global re3data.org portal for a fitting repository - this provides several filtering options.

It's not easy to evaluate the quality of repositories, because this is influenced by many external factors, starting with the mission of the repository. For instance, does it explicitly aim for long-term preservation - with the appropriate expertise and budget - or not? Is it dedicated to a specific research community and familiar with their data formats, or is it generic? However, if you focus on repositories that are certified as being trustworthy, you simplify your selection process. So, if you don't have a disciplinary repository, and use the re3data.org portal for your search, we recommend that you filter on "Certificate" and look for the red icon (unfortunately, OpenDOAR has no such filter).

SUPPORT ON METADATA, SENSITIVE DATA, USAGE LICENCES

Repositories may offer guidelines for sustainable data formats and metadata standards, as well as support for dealing with sensitive data and licensing. Most repositories provide persistent identifiers as well, so you may want to contact them early on.

You may notice that trustworthy repositories are sometimes stricter than non-certified repositories. This has to do with the fact that they need your help to carry out their long-term mission - good data management, which has many stakeholders, including data depositors themselves. For instance, repositories may ask researchers to provide so-called preferred file formats or documentation when they archive their data, along the lines of "FAIR data in trustworthy repositories". Read more about this in section "What are these repository certifications based on?".
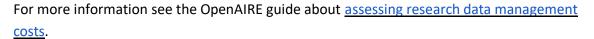
ARE DATA PUBLISHING COSTS SUPPORTED?

Many repositories don't charge you for depositing data with them (nor for downloading data, for that matter). The costs of making your data available via a repository are then limited to the time you may need to prepare the data for archiving, so it's recommended to plan this in a timely fashion.

In Horizon 2020 the costs related to making research data open access are eligible for reimbursement under the conditions defined in the Grant Agreement:

- if budgeted in the proposal and granted
- if claimed during the project

For more information see the OpenAIRE guide about assessing research data management costs.

WHAT ARE THESE REPOSITORY CERTIFICATIONS BASED ON? (BACKGROUND INFORMATION)

Here are the main requirements that a repository with a CoreTrustSeal complies with. Italics mark clear connections with the FAIRness of data. Remember: the Commission wants FAIR data in trustworthy repositories. Researchers and trustworthy repositories are partners in Open and FAIR data.

- Requirement 2 (R2): The repository maintains all applicable *licenses covering data access* and use and monitors compliance.
- R3: The repository has a continuity plan to ensure *ongoing access* to and preservation of its holdings.
- R4: The repository ensures, to the extent possible, that data are created, curated, accessed, and used *in compliance with disciplinary and ethical norm*s.
- R7: The repository guarantees the *integrity and authenticity of the data*.
- R8: The repository accepts *data and metadata* based on defined criteria to *ensure the relevance and understandability for data users*.
- R 10: The repository assumes responsibility for the *long-term* preservation and manages this function in a planned and *documented* way.
- R11: The repository has appropriate expertise to address *technical data and metadata quality* and ensures that *sufficient information* is available for end users to make quality-related evaluations.
- R13: The repository enables users to *discover the data* and *refer to them in a persistent way* through proper citation.
- R14: The repository enables reuse of the data over time, ensuring that *appropriate metadata* are available to support the understanding and use of the data.

HOW CAN OPENAIRE HELP?

Webpage https://www.openaire.eu/opendatapilot-repository-guide

Factsheet about the Open Data Pilot: https://www.openaire.eu/factsheet-h2020-odp

Information from the European Commission:
http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

Factsheet about Zenodo:

https://www.openaire.eu/factsheet-zenodo/view-document

Guide about identifying and assessing research data management costs:
https://www.openaire.eu/how-to-comply-to-h2020-mandates-rdm-costs

For data providers and repository managers OpenAIRE provides a validator to check if a given repository complies with the OpenAIRE guidelines: https://provide.openaire.eu/

VIEW OUR WEBINAR RECORDINGS

Tutorial "FAIR data in trustworthy repositories" (video):
https://www.youtube.com/watch?v=DutWdCYZ45I&list=PL0c4IRNUxuKe6DoqZfQEqlZ0eBjr9vvzu and slides: https://www.slideshare.net/OpenAIRE_eu/fair-ddata-in-trustworthy-repositories-the-basics