Data Monitor

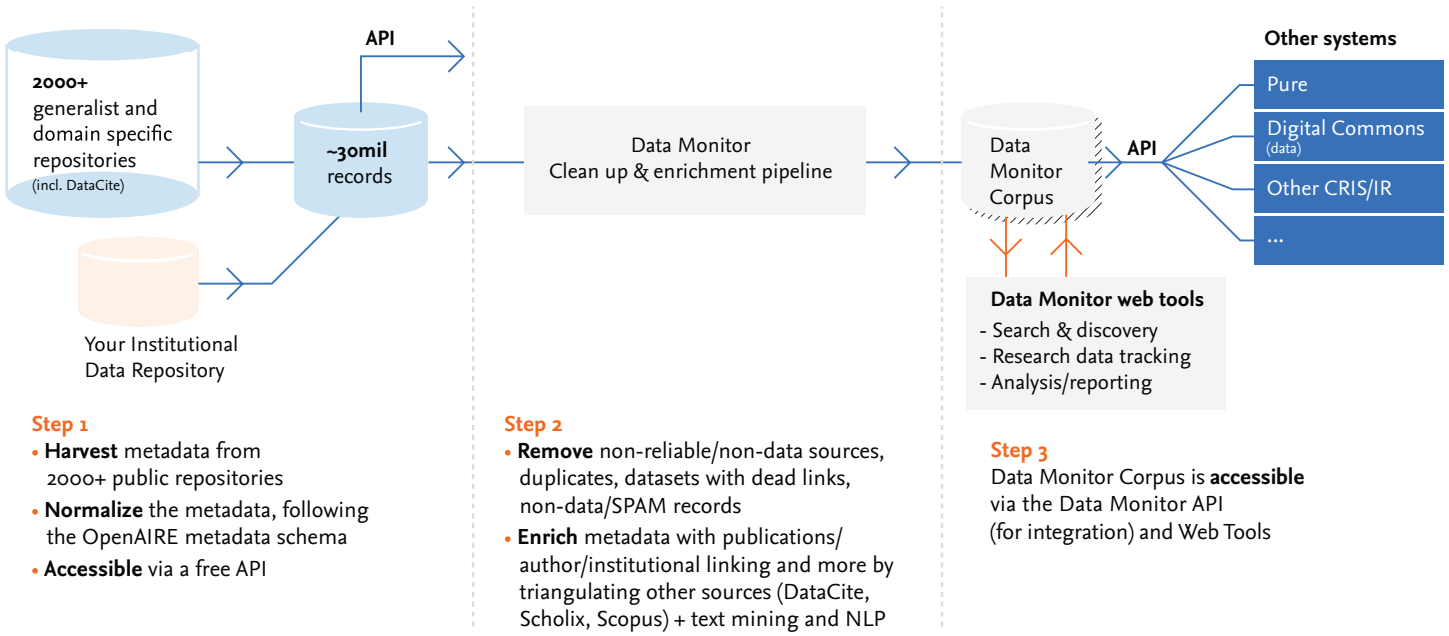
Content and Data Policy

ELSEVIER

TABLE OF CONTENTS

ELSEVIER

How Data Monitor works



Step 1
- **Harvest** metadata from 2000+ public repositories
- **Normalize** the metadata, following the OpenAIRE metadata schema
- **Accessible** via a free API

Step 2
- **Remove** non-reliable/non-data sources, duplicates, datasets with dead links, non-data/SPAM records
- **Enrich** metadata with publications/ author/institutional linking and more by triangulating other sources (DataCite, Scholix, Scopus) + text mining and NLP

Step 3
Data Monitor Corpus is **accessible** via the Data Monitor API (for integration) and Web Tools

Foreword

Data Monitor supports institutions with effectively tracking and analyzing the research data their researchers deposit across the various digital repositories available to them. To provide institutions the most accurate possible view on research data records affiliated with them, Data Monitor has created a corpus that is as comprehensive and as high-quality as possible: we strive to cover subject areas and geographies broadly and we have a few minimum metadata quality requirements. In the Introduction section, we define the scope of the repositories we index, which enable us to deliver a corpus with broad coverage. In the Content Selection section, we describe the several criteria we consider to effectively curate our corpus. We use these criteria as the basis to drive decision-making on which repositories we can index in the Data Monitor corpus. Note that, as both the research data management landscape and Data Monitor evolve, we will continue to revise and improve this policy to ensure that it remains relevant.

Introduction to Data Monitor

What is Data Monitor and what does it do?

Data Monitor is a tool geared towards librarians and research officers to track their institution's research data and monitor compliance with research data policies and mandates. Data Monitor harvests records from 2000+ repositories, either via direct ingestion or via DataCite, to create a corpus of research data records spanning multiple subject areas and geographies. All records are processed such that duplicates, spam or datasets with broken links are removed from the corpus. Data Monitor enriches the metadata of the harvested records with publications, author and institution links by triangulating sources such as DataCite, Scopus and ScholeXplorer. This reduces the efforts for institutions and repository managers to curate their research data metadata. For more details on how we build the Data Monitor corpus visit the product information page.

What are the sources indexed by Data Monitor?

Data Monitor indexes generalist, domain-specific, and institutional repositories directly or by harvesting the metadata records that these repositories make available through DataCite. These repositories are firstly assessed for scope and then on content and operational criteria as described in the following sections.

We strive to create a corpus that is comprehensive in terms of geographic and subject coverage. For this we do focus on repositories that are well known to their communities, of academic institutions or have already been appraised by stakeholders in the research data community such as by DataCite which indexes content from DataCite member repositories across all geographies and subject areas. To be indexed by DataCite, repositories must follow one of the supported metadata schemas.

Scope of the indexed repositories

Data Monitor focuses on the indexation of research data records hosted in digital repositories that provide public access to their holdings. When we refer to Research Data, we recognize the challenge in defining it and we therefore point to definitions recognized by the research community such as the definition coined by Christine Borgman (Borgman, 2015):

Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.

Or the definition included in the H2020 Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data:

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

Research data may be of various natures: primary/secondary data, raw/processed data, or discipline specific such as but not limited to metadata, molecular data, geospatial data, surveys, interviews, corpora of literature, images of artwork, etc. Repositories indexed by Data Monitor may contain other objects than research data (see Content selection criteria). As such, records indexed in our corpus may go beyond research data as defined above, and include software and code which we acknowledge are important in the obtaining or processing of research data. For a list of examples of the types of content that we consider in scope for indexation refer to the Annex Definitions of types of indexed content).

Out of scope for indexation in the Data Monitor corpus are objects building on the research data itself such as articles, thesis, reports, conference proceedings or other forms of publications *(table 2. in Annex Definitions of types of indexed content).*

Data Monitor focuses on establishing what objects are in scope for indexation and does not weigh on the underlying scientific quality of repository records.

Content selection criteria

Data Monitor focuses on repositories that contain research data. Whereas we aim not to index article-only repositories, we often index content from repositories that contain objects other than research data. These repositories are considered for harvesting on a case by case basis.

General criteria

Determining what constitutes a trustworthy data repository has been addressed by many stakeholders in the research data community and different approaches (certification, recommendation) have been suggested (Husen et al., 2017). Repositories can exhibit their trustworthiness through formal evaluation by certification bodies (e.g. CoreTrustSeal, nestor, ISO16363/TDR). While there is broad consensus on criteria to be met by repositories to be trustworthy (mission, access, infrastructure, etc.), recommendation is generally more widespread than certification, as it remains a challenge to apply certification standards to all existing repositories in use (Husen et al.).

At present, rather than prescribing specific criteria on trustworthiness, Data Monitor's policy encourages repositories to at the very least consider recommended best practices for digital repositories such as the TRUST principles. TRUST stands for Transparency, Responsibility, User focus, Sustainability and Technology, and according to these principles, repositories are responsible for providing reliable data services in a transparent manner and able to support long-term preservation of their research data holdings (Lin et al., 2020).

Operational criteria

Repositories will be assessed on their maturity status (e.g. scope, preservation and storage plan, documentation, etc.) as well as regarding access to the metadata of their records. These aspects are necessary to determine the scope and feasibility for Data Monitor to ingest the records of the candidate repository. We encourage repositories to provide their users clarity on their curation mechanisms used in alignment with the TRUST principles (Lin et al., 2020).

Metadata quality criteria

To ensure discoverability, research data should be described accurately and comprehensively with metadata in line with the FAIR principles for scientific data management (Wilkinson et al., 2016) We encourage repositories to align their schema with the Open AIRE metadata format which we align the metadata of records indexed in corpus with.

The Data Monitor team will evaluate repositories on:

- Metadata comprehensiveness: while this may vary among repository records metadata fields expected to describe a record are:
  - Author(s)
  - **Title**
  - **Persistent identifier (PID)**

Note that we are unable to harvest records that do not have a Title and PID.

- PID provenance: the PID of a research data record should be unique to that record and independent of the PID of other digital objects (i.e. the PID is not derived from the PID of an associated article). It is NOT a requirement that such PID is a DOI (digital object identifier).

Acknowledgements

We would like to thank the following research data community experts and academic institution representatives whose insights have greatly improved this document:

Jon Petters, Assistant Director of Data Management and Curation Services at Virginia Polytechnic Institute and State University (USA) and co-chair of the RDA/WDS Certification of Digital Repositories Interest Group; Madeleine de Smaele, Repository Manager of 4TU.ResearchData at the TU Delft Library; Tiina Sipola, Senior Information Specialist at University Library of Oulu, Finland; and finally the three co-chairs of the RDA Data Discovery Paradigms Interest Group Kathleen Gregory (Postdoctoral research fellow at the School of Information Studies, University of Ottawa, Ottawa), Fotis E. Psomopoulos (Principal Investigator at the Institute of Applied Biosciences at the Centre for Research and Technology Hellas, Thessaloniki, Greece) and Mingfang Wu (Senior Research Data Specialist at Australian Research Data Commons).

References

- Borgman, C.L. (2015). Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press.
- Husen, S.E., de Wilde, Z.G., de Waard, A. and Cousijn, H. Recommended versus Certified Repositories: Mind the Gap. Data Science Journal, 16, p.42 (2017). http://doi.org/10.5334/dsj-2017-042
- Lin, D., Crabtree, J., Dillo, I. et al. The TRUST Principles for digital repositories. Sci Data 7, 144 (2020). https://doi.org/10.1038/s41597-020-0486-7
- Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Annex

Definitions of types of indexed content

Types of content within the "Research Data" scope	
Term	**Definition***
Audio	A resource primarily intended to be heard.
Chemical Structure	A graphical representation of the arrangement of chemical bonds between atoms in a molecule (or in an ion or radical with multiple atoms)
Collection	An aggregation of resources that can contain multiple files
Computational Notebook	Notebook interface or a computational notebook is a virtual notebook environment used for literate programming (a method of writing computer programs).
Dataset	Data encoded in a defined structure
Event	A non-persistent, time-based occurrence. Metadata for an event provides descriptive information that is the basis for discovery of the purpose, location, duration, and responsible agents associated with an event (e.g exhibition, webcast, conference, battle, performance, etc.).
File set	Zipped files in different file formats (.rar , .zip, etc.)
Geospatial data	Spatial topic and spatial applicability may be a named place or a location specified by its geographic coordinates.
Image	A visual representation other than text.
Interactive resource	A resource requiring interaction from the user to be understood, executed, or experienced.
Model	Set of ideas and numbers that describe graphically or mathematically the past, present, or future state of empirical objects, phenomena, or physical processes.
Numerical data	Data expressed in numbers or a number system.
Physical Object	An inanimate, three-dimensional object or substance.
Service	A system that provides one or more functions
Sequencing Data	The exact order of bases in a nucleic acid or of amino acids in a protein.
Slide	An electronic image presented as a part of a series (e.g. PowerPoint slides).
Software	A computer program in source or compiled form
Tabular Data	Data arranged in a table (specifically: set up in rows and columns); data computed by means of a table.
Text	A resource consisting primarily of words for reading.
Video	A series of visual representations imparting an impression of motion when shown in succession. May or may not include sound.
Workflow	A sequence of tasks that processes data through a specific path

Types of content NOT in "Research Data" scope	
Term	**Definition***
Announcement	Public or formal words that announce an event or that constitute a public declaration or statement.
Article	A piece of writing about a particular subject that is included in scholarly journal, magazine, or newspaper.
Book	A set of typically printed pages that are held together inside a cover; a long written work contained in electronic or printed pages.
Conference paper	Article submitted for publication as contribution to a conference
Conference proceeding	Collection of scholarly articles published in the context of an academic conference or workshop.
Data paper	A data paper is a journal publication whose primary purpose is to describe data, rather than to report a research investigation. (Chavan et al., 2011)
Dissertation	An extended usually written treatment of a subject submitted for a doctorate.
Magazine	A periodical published online (containing miscellaneous pieces of writing and often illustrated).
Manuscript	Document submitted for publication (in an academic journal/book); includes article preprints.
Newsletter	A small publication containing news of interest (chiefly to a special group).
Report	A usually detailed written account or statement on a particular subject.

*Unless stated differently, the definitions listed in the tables above are quoted (in italic font) or adapted from the following sources: the Dublin Core Metadata Initiative terms and vocabulary; or the online dictionary Merriam Webster; or, if no other primary source available, from Wiktionary or Wikipedia.

Reference

- Chavan, V., Penev, L. The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics 12, S2 (2011). https://doi.org/10.1186/1471-2105-12-S15-S2

Document version: 1.0
Last updated: March 03, 2022