ARPHA Preprints

PENSOFT

PREPRINT

Author-formatted, not peer-reviewed document posted on 14/07/2023

DOI: https://doi.org/10.3897/arphapreprints.e109429

Current Developments in the Research Data Repository RADAR

Felix Bach, Kerstin Soltau, Sandra Göller, Christian Bonatto Minella, Stefan Hofmann

ARPHA Preprints *Author-formatted, not peer-reviewed document posted on 14/07/2023*. DOI: https://doi.org/10.3897/arphapreprints.e109429

Current Developments in the Research Data Repository RADAR

Felix Bach‡, Kerstin Soltau‡, Sandra Göller‡, Christian Bonatto Minella‡, Stefan Hofmann‡

‡ FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Eggenstein-Leopoldshafen, Germany

Corresponding author: Felix Bach (Felix.Bach@fiz-karlsruhe.de)

Reviewable v 1

Abstract

Effective research data management solutions and infrastructures are crucial for ensuring the long-term preservation, accessibility, and reusability of research data, not only to facilitate the verification and validation of discoveries but also to reduce or even avoid experimental redundancy. We present our first implementation steps of the Fair Digital Object Framework for RADAR, a generic research data repository, and discuss our experience with several approaches as well as their benefits and shortcomings. By embedding signposted typed link headers in our landing pages we increased the interoperability by making them machine-readable and -actionable, enabling the new data visiting paradigm additionally to metadata harvesting using the Open Archives Initiative Protocol for Metadata Harvesting.

Keywords

Research Data Management, FAIR, Repository, Digitalisation

Introduction

Effective research data management is crucial for ensuring the long-term preservation, accessibility, and reusability of research data, not only to facilitate the verification and validation of discoveries but also to reduce or even avoid the repetition of experiments (Wilkinson 2016). By making research data readily available to data users, potential redundancy in research can be minimised and the efficiency of funding improved, thereby enabling the data (re)use and accelerating scientific progress (Wilkinson 2016).

The provision of suitable infrastructure for archiving and disseminating research data is of paramount importance in this context, including the utilisation of data repositories (Wilkinson 2016).

One cross-disciplinary research data repository is RADAR hosted by Leibniz Institute for Information Infrastructure (FIZ) Karlsruhe since 2017. RADAR repository enables archiving and publishing service for any research data from both scientific studies and projects. It ensures access to and long-term availability of archived and published datasets supporting the FAIR principles (Findable, Accessible, Interoperable, and Reusable) (Wilkinson 2016).

The concepts, functions and features of RADAR are undoubtedly significant but may be optimised in the medium to long term, particularly in terms of interoperability.

The FAIR Digital Object (FDO) Framework appears to be a promising concept that can enhance data interoperability and address gaps in the current infrastructure and repository landscape (Schultes and Wittenburg 2018). This manuscript discusses RADAR's current development and implementation, focusing on FDO's potential for data repositories. We also examine how RADAR manages long-tail research data and its implications for research data management. In summary, this paper contributes to the understanding of the challenges and opportunities of cross-disciplinary research data management, interoperability and machine-actionability as well as the role of repositories like RADAR in addressing them.

RADAR - Research Data Repository

RADAR is a cross-disciplinary internet-based service for long-term and format-independent archiving and publishing of digital research data from scientific studies and projects. Its focus is on data from disciplines that are not yet supported by specific research data management infrastructures. The repository aims to ensure access and long-term availability of deposited datasets according to FAIR criteria (Wilkinson 2016) for the benefit of the scientific community. Published datasets are retained for at least 25 years; for archived datasets, the retention period can be flexibly selected up to 15 years. The RADAR Cloud service was developed as a cooperation project funded by the German Research Foundation (DFG) (2013-2016) and started operations in 2017. It is operated by FIZ Karlsruhe - Leibniz Institute for Information Infrastructure (re3data: http://doi.org/10.17616/R3ZX96, FAIRsharing: 10.25504/FAIRsharing.601a27).

The National Research Data Infrastructure (NFDI) that is currently set up in Germany, aims to create a community-driven sustainable digital infrastructure for research data management as an indispensable prerequisite for new research questions, findings and innovations. Within that framework, RADAR, originally designed for generic research data archiving and publication, now also addresses specific scientific communities - for example with subject specific metadata.

The new offerings RADAR4Chem and RADAR4Culture support the management and sharing of chemical research data, the first, and research data in the humanities and cultural sciences, the latter. Both data publication services are listed in re3data .

As a distributed, multilayer application, RADAR is structured into a multitude of services and interfaces. The system architecture is modular and consists of a user interface

(frontend), a management (backend) and a storage layer (archive), which communicate with each other via Application Programming Interfaces (API). This open structure and the access to the APIs from outside allows integrating RADAR into existing systems and work processes, e.g. for automated metadata (MD) upload from other applications using the RADAR API. RADAR's storage layer is encapsulated via the Data Center API. This approach guarantees independence from a specific storage technology and makes it possible to integrate alternative archives for the research data bitstream preservation (Fig. 1).

The data transfer to RADAR takes place in two steps: in the first step, the data is transferred to a temporary work storage. The ingest service accepts individual files and packed archives, optionally unpacks them while retaining the original directory structure. For each file found, the MIME Type (see Multipurpose Internet Mail Extensions specification) is analysed and stored in the technical MD. When archiving and publishing, in the second step, a dataset is created. The structure of this dataset - the AIP (Archival Information Package) in the sense of the OAIS standard - corresponds to the BagIt standard. In addition to the actual research data in original order, it contains technical and descriptive MD for the dataset (and optionally for each file or directory), as well as a manifest within one single TAR file ("tape archive", a unix archiving format and utility) as an entity in one place. The TAR file is stored permanently on magnetic tapes and redundantly as three copies at different locations in two academic computing centres.

Data harvesting, data visiting and FDOF

Traditionally, MD harvesting through the OAI-PMH protocol (Open Archives Initiative Protocol for Metadata Harvesting) involves an aggregator that collects (harvests) and stores MD from multiple sources in a standardised format. This allows users to search and discover resources across multiple repositories using a single interface. This approach is implemented in RADAR to allow integration with aggregators and central search services using our self developed massively scalable OAI-Provider software. However, this approach has several limitations. Firstly, it is a resource-intensive process that requires constant updating and maintaining the aggregated MD. In addition, the process of harvesting data can encounter various obstacles, ranging from inadequate technical infrastructure resulting in scalability issues to compatibility issues and varying standards, resulting in incomplete harvesting, misinterpreted or corrupted data. Finally, the approach is not very flexible, as it relies on a static central repository that cannot easily adapt to changes in MD and data.

Recently, especially because of diseases outbreak and above all the COVID-19 pandemic (Basajja et al. 2022), the traditional approach of MD harvesting through the OAI-PMH protocol and storing it in a central place where it is indexed and searchable, was challenged by the new concept of data visiting. The approach enables algorithms to automatically access and query multiple datasets while keeping the data within the institution where it originated.

This section presents a discussion on the benefits of data visiting, particularly in the context of the FDO Framework (FDOF) and how this can potentially revive the idea of the semantic web.

FDOF defines a model to represent objects in a digital environment and a mechanism to create, maintain and (re)use these objects. (Schultes and Wittenburg 2018). The framework consists of a predictable identifier resolution behaviour, a mechanism to retrieve the object, its MD and an object typing system. The main goal of the FDOF is to define a set of features to provide foundational support for the FAIR principles. One very important principle is the interoperability ('I'), that includes machine-readability and actionability of (mostly digital) objects.

By targeting the datasets' landing pages directly, machines can access both MD and data dynamically, since no previous data aggregation is required. Moreover, data visiting enables accessing MD (in a standardised format) and data as the latest version. Finally a greater automation and machine-actionability of research data, is promoted.

The FDO Framework, coupled with the concept of data visiting, has the potential to bring back the idea of the semantic web, albeit the idea was never fully implemented due to technical limitations and a lack of standardisation.

Implementing the FAIR Digital Object Framework (FDOF)

Implementation of landing pages

Current data repositories (such as RADAR) offer HTML-based landing pages for datasets that are optimised for human readability. They cannot be visited by "machines" to find links to MD, data and licence information. Landing pages are created for each RADAR dataset - both published or archived - in order to provide humans the information they require. Data publications´ landing pages are accessed through DOIs (Digital Object Identifiers) and offer descriptive MD in structured text form. These MDs are also available as downloadable files in different formats (RADAR-format, DataCite, OAI-DC, PREMIS, BibTex, EndNote, RIS, Fig. 2). Further, users can also download the whole dataset (as a TAR file) which includes a BagIt container with both data and MD.

Although this design may be optimal for human use, it presents a challenge for machine-assisted long-term archiving as the landing pages are optimised to cater to humans with prominent download buttons, making it difficult for machines to locate requested information by following signposted links or crawling through the HTML code. With dozens of hyperlinks embedded in the landing pages, it is difficult for machines to locate the link to the actual research data. Further, currently, machines have no mechanism to control licence information or directly read MD-fields to determine the relevance of a dataset to a semantic query without downloading and parsing the entire package.

In order to tackle these challenges and enhance the ability of machines to read and act upon the information, we have examined various approaches, which will be discussed in the subsequent sections.

Implementation of FAIR signposting in landing pages

The signposting community aims to increase the machine friendliness of scientific webpages. FAIR Signposting is currently not a formal standardisation process but follows a low-threshold approach that enables machines to interact with science portals in a uniform way, bypassing proprietary APIs. We discussed this approach at the 1st International Conference on FAIR Digital Objects (October, 26th-28th, 2022 in Leiden, the Netherlands) where it was regarded as promising to significantly promote the interoperability of scientific objects on data repositories and publishing platforms and thus implementing the FAIR Digital Objects (FDO) specification in HTML-based landing pages research data repositories.

The FAIR signposting concept is entirely based on widely used web protocols specified in I ETF RFCs. Patterns that occur repeatedly in scientific portals (e.g. DOI, author PIDs, web addresses of landing pages and descriptive MD) are clarified via the inclusion of typed links (FAIR Signposting Profile). These are provided in HTML resources in simple HTTP link headers and HTML link elements. Machine agents thus receive uniform and lightweight information about both elements and their web addresses and can navigate to them.

In line with the FAIR signposting approach, we enriched the source code of all landing pages of RADAR datasets with typed links, thus improving their machine readability and actionability in a state-of-the-art way. Following output of the curl command requesting the header of a dataset landing page shows the included typed links:

curl -I 'https://www.radar-service.eu/radar/de/dataset/ckPELTYwdlqHnNxe'HTTP/1.1 200**Date**: Mon, 26 Jun 2023 07:53:33 GMT**Server**: Apache/2.4.6 (CentOS) OpenSSL/1.0.2k-fips**Strict-Transport-Security**: max-age=31536000; includeSubDomains**Vary**: Origin,Access-Control-Request-Method,Access-Control-Request-Headers**X-Frame-Options**: DENY**X-Content-Type-Options**: nosniff**X-XSS-Protection**: 1; mode=block**Strict-Transport-Security**: max-age=31536000; includeSubDomains**Content-Security-Policy**: default-src 'none'; script-src 'self' https://*.radar-service.eu:* http://*.fiz-karlsruhe.de:* https://*.fiz-karlsruhe.de:* https://*.orcid.org https://*.lobid.org https://lobid.org 'unsafe-inline' 'unsafe-eval'; object-src 'self'; style-src 'self' https://*.radar-service.eu:* http://*.fiz-karlsruhe.de:* https://*.fiz-karlsruhe.de:* 'unsafe-inline'; img-src 'self' https://*.radar-service.eu:* http://*.fiz-karlsruhe.de:* https://*.fiz-karlsruhe.de:* https://maps.googleapis.com https://assets.crossref.org https://orcid.org https://lobid.org data: *.maps.deutsche-digitale-bibliothek.de http://localhost:*; media-src 'self' https://*.radar-service.eu:* http://*.fiz-karlsruhe.de:* https://*.fiz-karlsruhe.de:*; font-src 'self' https://*.radar-service.eu:* http://*.fiz-karlsruhe.de:* https://*.fiz-karlsruhe.de:*; connect-src 'self' https://*.radar-service.eu:* http://*.fiz-karlsruhe.de:* https://*.fiz-karlsruhe.de:* http://localhost:* https://*.orcid.org https://lobid.org https://*.lobid.org https://api.crossref.org https://api.ror.org; child-src 'self' https://*.radar-service.eu:* http://*.fiz-karlsruhe.de:* https://*.fiz-karlsruhe.de:***Link**: <https://doi.org/10.35097/946> ; rel="cite-as"**Link**: <https://creativecommons.org/licenses/by/4.0/deed.de> ; rel="license"**Link**: <https://orcid.org/0000-0002-7148-7210> ; rel="author"**Link**: <https://www.radar-service.eu/radar-backend/archives/ckPELTYwdlqHnNxe/versions/1/content> ; rel="item" ; type="application/x-tar"**Link**: <https://www.radar-service.eu/radar/de/export/ckPELTYwdlqHnNxe/exportdatacite> ; rel="describedby" ; type="application/vnd.datacite.datacite+xml"**Link**: <https://www.radar-service.eu/

radar/de/export/ckPELTYwdlqHnNxe/exportbibtex> ; rel="describedby" ; type="application/x-bibtex"**Link**: <https://www.radar-service.eu/radar/de/export/ckPELTYwdlqHnNxe/exportradarmetadata> ; rel="describedby" ; type="text/xml" **Content-Type**: text/html;charset=UTF-8**Content-Language**: de**Transfer-Encoding**: chunked**Set-Cookie**: JSESSIONID=6AEE132A9E1AE119400DA9A1EE6923ED; Path=/radar; HttpOnly**Set-Cookie**: org.springframework.web.servlet.i18n.CookieLocaleResolver.LOCALE=de; Path=/; HttpOnly

Discussion on embedding metadata like schema.org in the landing page HTML

While the concept of signposting involves providing explicit and standardised links to the dataset and its attributes, embedding MD in landing pages involves including machine-readable MD in a format such as JSON-LD or RDF. The latter enables web crawlers (such as Google) to find typed entities in the embedded MD and include them in a search index. Standard vocabularies used for adding such structured data to web pages, such as Schema.org , make it easier for search engines to understand and categorise the content using a number of different types of metadata markup. This concept, at a first glance, seems to be a promising approach enabling search engine crawlers to better interpret the meaning of landing pages for datasets. Some data repositories already have adopted this approach for making their datasets more discoverable on the web, including NASA, NOAA and Harvard's Dataverse. However, this approach has a major shortcoming namely that machines, like crawlers, need to parse the HTML and find and interpret the embedded metadata, which is slow and not optimised for live queries.

Discussion of the different approaches and their implementation

Both techniques (signposting and embedding MD in landing pages of datasets) enhance machine-readability and -actionability of research data in the repository, allowing for more efficient and effective research.

In the signposting approach however, machines do not need to parse the landing pages, resulting in faster access to relevant information. We decided to provide this advantage for machines and implemented signposting in our RADAR landing pages.

Furthermore, we embedded schema.org MD in the HTML of dataset landingpages because this offers possibilities to include MD (also discipline specific MD, e.g. bioschema.org) resulting in better findability for users searching in standard search services that can crawl the webpages and process embedded MD. This approach adheres to an open standard and represents a shift towards machine-based data findability.

By adopting the FDOF in such a lightweight way, repositories like RADAR can promote new possibilities for data sharing, (re)using, and thus enabling interoperability across different research domains by increasing the interpretability of their landing pages.

Following this approach, other data repositories and scientific communities could considerably benefit from adapting repository landing pages, such as in the RADAR example, with low effort and no radical change in their architecture or processes (Fig. 3).

Summary

We presented our first implementation steps of the FDOF in RADAR, a generic research data repository, with the goal to increase the FAIRness of included datasets. We discussed our experience with several approaches and their benefits as well as shortcomings.

By embedding signposted typed links in the headers of landing pages that correspond to datasets, we made it much easier for machines to directly find a specific part of information like metadata attributes and their type and to download and process the dataset, knowing its type in advance. This makes RADAR and its datasets more interoperable with other data services by increasing machine-actionability. By additionally enriching the landingpages with schema.org annotation, we made it easier for web crawlers to extract descriptive metadata in a standardised way, leading to a better findability in overarching search services such as the google dataset search. Making datasets machine-readable and -actionable, enables the new paradigm of data visiting additionally to MD harvesting using OAI-PMH.

Acknowledgements

We would like to express our gratitude to the National Research Data Infrastructure (NFDI) as well as the NFDI4Chem and the NFDI4Culture consortia.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Basajja M, Suchanek M, Taye GT, Amare SY, Nambobi M, Folorunso S, Plug R, Oladipo F, Reisen M (2022) Proof of Concept and Horizons on Deployment of FAIR Data Points in the COVID-19 Pandemic. Data Intelligence 4 (4): 917-937. https://doi.org/10.1162/dint_a_00179
- Schultes E, Wittenburg P (2018) FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure. *Communications in Computer and Information Science* URL: https://link.springer.com/book/10.1007/978-3-030-23584-0
- Wilkinson M, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data https://doi.org/10.1038/sdata.2016.18

Author-formatted, not peer-reviewed document posted on 14/07/2023. DOI:
https://doi.org/10.3897/arphapreprints.e109429

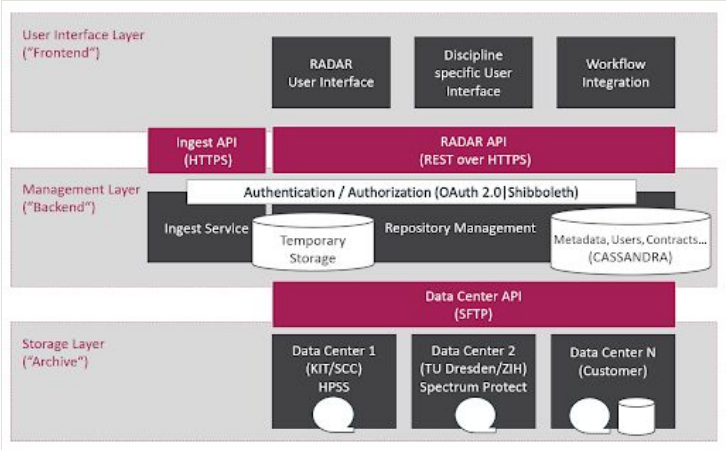

Figure 1.

Overview of RADAR´s architecture: RADAR offers a graphical user interface that is connected to the management layer by an API. Other user interfaces can easily be added. At the same time, several storage backends at different data centres can be connected to the backend part of the management layer by an API.

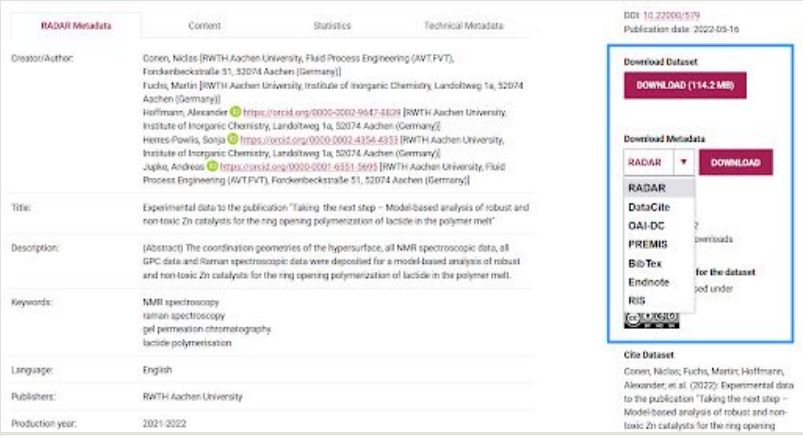

Figure 2.

Example of the download options of dataset and metadata in several formats, embedded in a RADAR landing page.

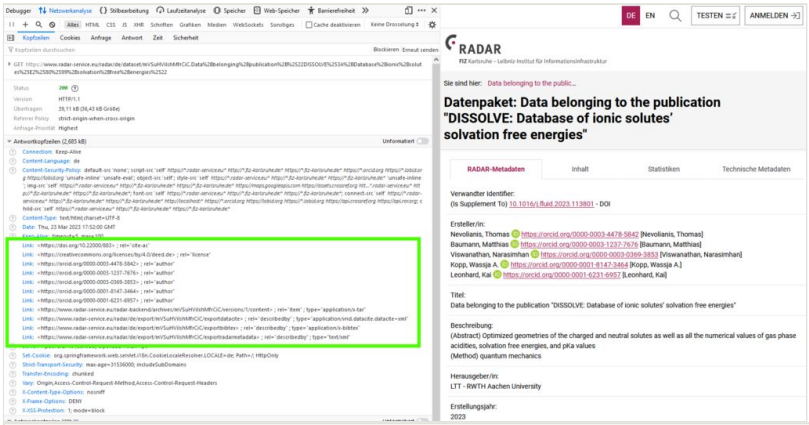

Figure 3.

RADAR´s source code of a dataset landing page including typed links.