RADAR: A Research Data Management Repository for Long Tail Data

Ena Brophy[1], Matthias Razum[2]

1,2 FIZ Karlsruhe – Leibniz Institute for Information Infrastructure

Abstract. The transparency and reproducibility of scientific results are increasingly based on digital data. In compliance with good scientific practice, data needs to be published, accessible, and re-usable. RADAR is a generic infrastructure providing archival and publication services for research data. RADAR focuses on the "long tail of science", which often lacks sufficient research data infrastructure. It offers two service levels: data archival ("dark archive" with variable retention periods and user-defined access rights) and data archival with publication (guaranteed retention period of 25+ years, DOI assignment, and flexible licensing options).

Users can upload, edit, structure and describe (collaborative) data in an organisational workspace. Administrators and curators can manage access and editorial rights before the data enters the preservation and optional publication level. Data consumers may search, access, download and retrieve usage statistics on the data via the RADAR portal. For data consumers, findability of research data is of utmost importance. The metadata of published datasets can be harvested via a local OAI provider or the DataCite Metadata Store. Additionally, RADAR provides an application programming interface (API) for easy integration of RADAR functionality with existing systems and workflows at the user's side.

RADAR relies on two academic data centres under German jurisdiction to provide its services. The novel two-stage service and business model provides one-off payments and institutional subscription services. RADAR is intended to become an integral part of the international information infrastructure which also allows the integration of third-party services. RADAR was designed by a research consortium of academic institutions for the academic community. Through cooperation with researchers, data centres, scientific societies as well as publishers, RADAR ensures that the resulting infrastructure is designed to meet the requirements of the academic community.

Keywords: RADAR, research data repository; repository; preservation; information infrastructure; research data management; data archiving; data publication.

Introduction: Data Management for the Long Tail

The collection and organisation of data is a fundamental element of the research process. In compliance with good scientific practice, data needs to be published, accessible, and re-usable. Digital data offer the potential for greater return on investment, provided that data is properly managed and shared among researchers (Berman, et al. 2010, Buckland 2011). The academic community is becoming more interested in collecting and providing access to datasets produced at their institution for reuse. Driving this is the transparency and reproducibility of scientific results which is recognised as a primary research output based on digital data (Treloar und Harboe-Ree 2008, Klump 2009, Neuroth, et al. 2012). While the focus has been on the accessibility of 'big data', i.e. disciplines whose output produces large volumes of data, many research studies produce smaller

datasets. This poses a challenge to the academic community who needs to manage and sustain access to research data that does not necessarily fall within the scope of discipline-based solutions. A survey conducted by the journal Science in 2011 stated that 48.3% of respondents were working with datasets that were less than 1GB in size, and over half of respondents reported that they stored their data only in their laboratories (Science Editorial 2011). Solutions may differ from discipline to discipline in size, scale, project duration etc. (C. L. Borgman 2015, Borgman, et al. 2015). This is true in particular for long tail data which often lacks sufficient research data infrastructure. Best practice for the data management of long tail data is often dependent on the community.

What has emerged in the last number of years for both big and long-tail data is the need for being open to be searched, cited and downloaded for potential reuse. Funders such as the National Science Foundation require researchers to include a data management plan as part of their proposal for funding (National Science Foundation 2011). In Germany, the German Research Foundation published guidelines for "Safeguarding Good Scientific Practice" to ensure that data produced as part of scientific studies are recognised as primary research output (DFG 2013). In 2016, stakeholders from academia, industry, publishers and funding agencies published a concise and measurable set of principles called the FAIR Data Principles (Findable, Accessible, Interoperable and Re-usable). These principles place specific emphasis on enhancing the ability of machines to automatically find and reuse data (Wilkinson 2016, FORCE11 2016). To highlight the importance of keeping data FAIR, the European Commission adopted the FAIR Data Principles and released new Guidelines on FAIR Data Management in Horizon 2020 (European Commission 2016). The EC guidelines include several important changes that aim to improve the quality of project results, achieve greater efficiency, and achieve progress and growth of a transparent scientific process. Consequently, research institutions, universities and libraries are becoming more interested in collecting and providing access to datasets produced at their institution that do not fall within the scope of big data or discipline-based repositories. In addition, researchers themselves start to look for data services. This paper presents a multidisciplinary solution - the RADAR (Research Data Repository) service[1], a generic research data repository for data preservation and publication in research to include the social sciences and humanities.

The RADAR Service

RADAR was developed as a cooperation project of five research institutes from the fields of natural and information sciences[2]. The technical RADAR infrastructure is provided by the FIZ Karlsruhe – Leibniz Institute for Information Infrastructure and the Steinbuch Centre for Computing (SCC), Karlsruhe Institute of Technology (KIT). The sustainable management and publication of research data with DOI-assignment was provided by the German National Library of Science and Technology (TIB). The Ludwig-Maximilians-Universität Munich (LMU), Faculty for Chemistry and Pharmacy and the Leibniz Institute of Plant Biochemistry (IPB) provided the scientific knowledge and specifications and ensure that RADAR services can be implemented to become part of the scientific workflow of academic institutions and universities.

1 RADAR (Research Data Repository): https://www.radar-service.eu
2 RADAR Project website: https://www.radar-projekt.org/display/RD/Home

The heterogeneity of research data is an important issue for the research community. Therefore the primary goal of RADAR was to establish an interdisciplinary research data repository, which is sustained by research communities and supported by a stable business model. RADAR has approached this problem by focusing on real scientific workflows and established best practice throughout the testing phase of product development. During the project phase, a number of public workshops were held to gather requirements and discuss technical, organisational and legal matters. Furthermore, a scientific advisory board was established. A wealth of requirements, feedback, and advice was collected via both channels. After the first two years of the project a test system was launched enabling the academic community to test RADAR using datasets from different subject areas. This approach provided the project team with significant insight, which allowed them to design the service with the academic community at its centre to ensure the service will be effective.

The upload of scientific data into repository collections is a continuing challenge for researchers and their affiliated research institutions. To facilitate this RADAR provides a generic infrastructure, which delivers archival and publication services for research data. RADAR offers a suite of services to ensure that the requirements of funding agencies and good scientific practice are met.

Basic service: Data Preservation

For data providers, RADAR offers format-independent preservation services to store data in compliance with specific institutional or funder requirements periods (e.g., 10 years according to DFG recommendations). This includes secure preservation of up to 15 years with the data remaining unpublished, and the requirement of a minimum set of metadata. By default, the data and associated metadata will not be published, unless specified otherwise by the data provider. RADAR offers a flexible data and metadata access management so that data providers are able to share preserved datasets with other RADAR users if desired and manage the external visibility of the associated metadata.

Extended Service: Data Publication

For making data citable, traceable and reusable, RADAR offers a combined service of research data archival and publication. Datasets published in RADAR are identified by DOI. Using the DOI, datasets can be referenced persistently and unambiguously. The service also includes an optional embargo period for the publication of submitted data that may be subsequently prolonged if necessary. The metadata describing the dataset is published already during the embargo and datasets are allocated a DOI. This ensures that datasets can be found and cited already when they are deposited, while downloads will only be possible once the embargo period has expired. Within the publication service, a peer review option may be used: In this case, the respective dataset is "frozen" for the duration of the peer review process and receives a secure "review-URL" provided by RADAR which may be forwarded to an editor or reviewer responsible for a corresponding manuscript submission. As such, manuscript and data may be inspected simultaneously during a review process.

Architecture

The system architecture is based on an expendable API structure, referred to as 'Computing Centre API' in Fig. 1. This structure allows an integration of multiple computing centres that use various storage systems (e.g. TSM, SamQFS, DMS, HPSS). To reach a uniform archiving interface, the API hides these various storage systems and technologies. The storage is managed by using a repository software which consists of two parts. A back end addresses general tasks such as storage access and bitstream preservation, whereas the front end implements RADAR-specific workflows. Front-end workflows include various data services: Metadata management, access control, data ingest processes, as well as the licensing for reuse and publishing of research data with DOI. Archival Information Packages and Dissemination Information Packages are provided in a BagIt-structure in ZIP format. The RADAR API enables users to integrate the archival backend into their own systems and workflows. RADAR stores the data in two academic data centres under German jurisdiction. The Steinbuch Centre for Computing (SCC) at Karlsruhe Institute of Technology (KIT) acts as the primary data centre, holding two copies of the data at two different locations. A third copy is replicated to the Zentrum für Informationsdienste und Hochleistungsrechnen at TU Dresden. The metadata catalogue and the software is hosted by SCC. The two data centres employ different hardware and software systems as well as differing administrative procedures. This approach adds an additional level of security and helps avoiding systematic errors that may put corrupt large chunks of the archived data.

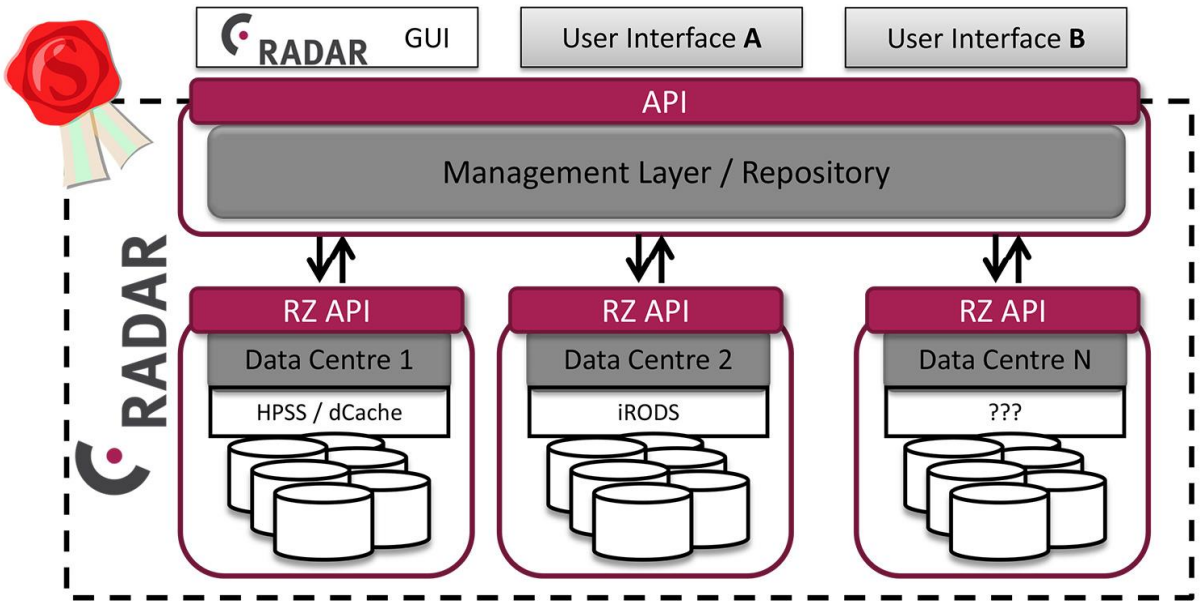

Figure 1. RADAR architecture with data ingest and API (Brophy & Razum 2017)

TIB Hannover provides as German DataCite agency the necessary systems and interfaces for registering DOIs assigned to published data sets. All communication between RADAR and DataCite is handled via REST calls.

Metadata Schema

Metadata are essential to the traceability, access and effective use of scientific data. In RADAR, submitted data must be accompanied by a set of basic descriptive metadata elements that document and describe a particular resource. The following scheme aims to enhance the traceability and usability of research data by maintaining a discipline agnostic character and simultaneously allowing a description of discipline specific data.

The RADAR Metadata Schema (Table 1.) includes ten mandatory fields, which represent the general core of the schema[3]. These fields contain the main requirements for DOI registration, in accordance with DataCite Metadata Schema 4.0. (DataCite 2016) and must be supplied when submitting metadata to RADAR. Additionally, 13 optional metadata parameters serve the purpose of describing discipline specific data.

Table 1. RADAR Descriptive Metadata Schema. The Schema contains a mandatory set of generic parameters to allow for the accurate and consistent identification of a resource for citation and retrieval purposes. Subsequently, optional parameters can be described, to meet the requirements of discipline-specific datasets.

Descriptive Metadata: 10 standard parameters	Mandatory parameters for basic information
Identifier (RADAR ID/DOI)	A unique string, which identifies a resource. Handle for data preservation/DOI for Data publication service)
Creator	Persons involved in producing the data
Title	Study/Data title
Publisher	Corporate/Institutional or personal name
Production Year or time span	Year, in which data was created or refers to
Publication Year	Year, in which the resource was published
Subject Area	Scientific fields appropriate for the resource
Resource	Resources content (e.g. dataset, model, software)
Rights	Rights management statement (e.g. CC BY)
Rights holder	Institution/Person holding rights.
Descriptive Metadata: 13 optional parameters	**Parameters for discipline specific data description**
Additional title	Additional title type (e.g. translated title)
Description	Further information (e.g. abstract)
Keyword	Keywords describing the subject focus
Contributor	Associated institution/person
Language	Primary language used or relevant to resource
Alternative Identifier	Unique string within its domain of issue (e.g. local identifier)
Related Identifier	Identifiers of related resources
Geo Location	Region/Place where resource originated/refers to
Data Source	Data origin (e.g. instrument, observation, trial)
Software type	Software used for data production and processing
Data Processing	Specifies further processing (e.g. statistics)
Related Information	Further information (e.g. database number)
Funder Information	Funder information

The parameters were implemented with a combination of controlled vocabularies and free-text entries, thereby covering heterogeneous data produced by a multiple of disciplines. The controlled

3 RADAR Schema documentation: https://www.radar-service.eu/en/radar-schema

vocabulary entries were defined in accordance with established regulations in mind (e.g. ISO standards). RADAR clients, who wish to enhance the prospects of their metadata being found, cited and linked to original research are strongly encouraged to submit the optional parameters in addition to the mandatory set of properties. The metadata of datasets that are published in RADAR will be available under the Creative Commons Zero licence (Creative Commons 2014) RADAR will actively disseminate all published metadata to DataCite. Metadata of datasets that are only archived (not published) in RADAR are only available to the data provider, unless otherwise specified. Moreover, a support service for data harvesting of published metadata via OAI-PMH interface is provided.

Business Model

The business model, including the services presented in the previous section, ensures a sustainable operation environment for the data archive as well as for institutional users. From the start, RADAR focuses on publicly funded research institutions and universities in Germany. This limitation is mainly driven by contractual and legal issues. RADAR strives to loosen some of the limitations in the near future to broaden the potential user base and expand to neighbouring European countries.

The ongoing operation of RADAR is not based on project funding. Operational costs include personnel, marketing and travel expenses and fees for the basic IT infrastructure. Half of these costs are taken over by FIZ Karlsruhe, which understands RADAR as an important building block of the information infrastructure and an excellent fit for its mission. The other half of the operational costs and all variable costs (which are mainly the costs for maintaining three copies of the data in two data centres) are factored into the pricing of the service. Being charged for such a service might turn away researchers, but at the same time, it might be a healthy exercise to re-evaluate the data produced in the course of a project and decide which data needs to be published, which can be archived and which might even be deleted.

RADAR offers two different pricing models for the two service levels: for archived data, the amount of stored data defines the price per year[4]. Institutions may end contracts and move the data to other service providers any time. For published data, the message from the academic community was very clear that there needs to be a guarantee that the data is available independent from the contractual situation. Thus, RADAR offers a one-time payment model for published data with a guaranteed retention period of at least 25 years. One-time payments also work well with the research system which relies in most cases on project funding with no option for ongoing payments after the end of a project. Due to the corporation with outstanding partners, RADAR can offer very competitive pricing for its service[5].

Conclusion & Outlook

As universities and research institutions are increasingly interested in collecting and providing access to datasets produced at their institution, not all of this data will fall within the scope of big

4 RADAR Pricing Information: https://www.radar-service.eu/en/pricing
5 RADAR Pricing Structure: https://www.radar-service.eu/en/pricing

data or discipline based repositories. The researchers from long tail of science will start to look to libraries to provide support and data services for their datasets.

With RADAR, we present a solution that has been designed by a research consortium of academic institutions for the academic community. This interdisciplinary approach, competitive pricing, option to integrate RADAR with existing services and workflows, and compliance with German and European legislation makes RADAR a viable option for research data archival and publication. The novel two-stage service and business model combined with a trustworthy repository for institutions and their researchers provides a contribution to ensure a better availability, sustainable preservation and publishability of research data for present and future academic communities.

Acknowledgements

We gratefully acknowledge the tremendous efforts of everyone involved in providing support throughout the project. In particular, we wish to thank the RADAR project team for their ongoing input to the RADAR service. We thank the scientific advisory board of RADAR for their contributions to our discussions on data management, research data services and infrastructures. We also thank the academic community for evaluating the test system. Their constructive comments have helped to improve the RADAR service before we moved to production.

Funding

RADAR was developed as part of a three-year project funded by the German Research Foundation (DFG) from 2013 to 2016 (http://www.radar-projekt.org) and is placed within the program "Scientific Library Services and Information Systems (LIS)" on restructuring the national information services in Germany.

References

Berman, Francince, Brian Lavoie, Paul Ayris, G. Sayeed Choudhurry, Elizabeth Cohen, Pual Courrant, Lee Dirks, Amy Friedlander, Vijay Gurbaxani, Anita Jones, Ann Kerr, Clifford Lynch, Daniel Rubinfeld, Chris Rusbridge, Roger Schonfeld, Abby Smith Rumsey, and Anne Van Camp. 2010. *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information: Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access.*

Borgman, Christine L., Peter T. Darch, Ashley E. Sands, Irene V. Pasquetto, Milena S. Golshan, Jilian C. Wallis, and Sahron Traweek. 2015. "Knowledge infrastructures in science: Data, diversity, and digital libraries." *International Journal on Digital Libraries* 16, 3: 207–227.

Borgman, Christine L. 2015. *Big data, little data, no data: Scholarship in the networked world*. Cambridge, MA: MIT Press.

Buckland, M. 2011. “Data management as bibliography.” *Bulletin of the American Society for Information Science and Technology* 37: 34–37.

Creative Commons. 2014. Creative Commons Licenses. https://creativecommons.org/licenses/.

DataCite. 2016. DataCite Metadata Schema 4.0. https://schema.datacite.org/.

DFG, Deutsche Forschungsgemeinschaft. 2013. Safeguarding Good Scientific Practice. Recommendations of the Commission on Professional Self Regulation in Science.: WileyVCH.

European Commission. 2016. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

FORCE11. 2016. https://www.force11.org/group/fairgroup/fairprinciples.

Klump, J. 2009. “Managing the Data Continuum.” http://oa.helmholtz.de/fileadmin/user_upload/Data_Continuum/klump.pdf.

National Science Foundation. 2011. “Proposal Preparation Instructions.” Grant Proposal Guide.

Neuroth, H, S Strathmann, A Oßwald, R Scheffel, J Klump, and J Ludwig. 2012. “Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme.”
Science Editorial, 2011. “Challenges and Opportunties.” Science 331. 6018: 692-693.

Treloar, Andrew, and Cathrine Harboe-Ree. 2008. “Data management and the curation continuum: how the Monash experience is informing repository relationships.”

Wilkinson, Mark D. et al. 2016. “The FAIR Guiding Principles for scientific data management and stewardship.” Nature, March.