Proceedings of the IATUL Conferences

2015 IATUL Proceedings

Jul 6th, 12:00 AM

RADAR - A repository for long tail data

Angelina Kraft
TIB, Hanover, Germany

RADAR - A REPOSITORY FOR LONG TAIL DATA

Angelina Kraft[1*], **Janna Neumann**[1]
[1]German National Library of Science and Technology TIB, Hannover, Germany
[*]Corresponding author: angelina.kraft@tib.uni-hannover.de
janna.neumann@tib.uni-hannover.de

Abstract

The way knowledge is shared is experiencing a paradigm shift: Digital networks allow new degrees of openness for research and its resources, accompanied by a huge potential for scientists, inventors, industry and the general public. Accessible data will allow all groups to participate in innovation and value creation regardless of their geographical location or individual background. However, for researchers who are evaluated by their academic performance and scientific excellence, there is a fine balance between benefits and concerns regarding the openness of resources such as knowledge and data. With the Research Data Repository (RADAR) project we provide solutions to maintain this balance: In RADAR, an interdisciplinary infrastructure for the preservation, publication, creditability and traceability of research data from the fields of the 'long tail of science' is developed.

Here we present the first RADAR prototype: A robust, generic end-point data repository which enables clients to preserve research results up to 15 years and assign well-graded access rights, or to publish and preserve data with a DOI assignment for an unlimited period of time.

Potential clients include libraries, research institutions, publishers and open platforms which require an adaptable digital infrastructure to archive and publish data according to their institutional needs and workflows.

In a nutshell, RADAR can help clients to handle following issues:

- Secure storage of research data.
- Preservation of information after a project is completed, a grant ends or employees leave.
- Traceable and citable data publication across communities via a discipline-agnostic metadata scheme.
- Ensuring that data are ‘stable’ after publication e.g. to allow accurate comparisons later.
- Provision of data management services for their customers up front while using RADAR as a back-end system.

Keywords: repository, preservation, information infrastructure, research data management, data storage

1 Libraries & digital data production: New infrastructures & future benefits

Nowadays, data underlying scientific studies starts to be recognized as a primary research output, complementing widely acknowledged research publications such as journal articles and books (Treloar & Harboe-Ree 2008, Klump 2009, Neuroth et al. 2012). Facing the rapid increase in digital data production, the world-wide challenges to curate, publish and sustainably access research data are enormous. Until now, much focus has been given to improve the accessibility of so-called ‘big data’. Big data include extensive datasets produced through large science projects, such as particle physics research carried out at CERN (e.g. using the Large Hadron Collider) (Whyte et al. 2015). However, there are thousands of research studies that produce ‘smaller’ datasets. In 2011, a survey of 1700 researchers across disciplines was undertaken by the journal *Science*. They found that 48.3% of respondents were working with datasets less than 1GB in size, and over half of those polled store their data only in their laboratories (Science 2011). Such heterogeneous data collections occur across various scientific disciplines and are called the ‘long tail of research data’. Long tail disciplines are characterized by hypothesis-driven studies led by small investigation groups who produce and analyze their own datasets (Harris 2011, Thessen & Patterson 2011). Moreover, the use of standards and best practices for data management is extremely variable in the long tail and datasets often lack a concise structure. The survey by Science (2011) concluded that “(…) *different institutions archive their research data in different ways - making access difficult from* *outside the institution*". For research communities within Europe, a similar picture is found: The framework behind the HORIZON 2020 program stated that "(…) *diversity is likely to remain a dominant feature of research data – diversity of formats, types, vocabularies, and computational requirements – but also of the people and communities that generate and use the data*" (European Commission 2013).

Consequently, universities and research institutions 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. Furthermore, researchers themselves start to look for services which facilitate data management processes. These conditions bring new opportunities for libraries to provide support and data services, e.g. by co-operating with data centers and research institutes. As a result of such co-operations, libraries can offer generic services that also meet the highly subject-specific data management requirements of specialized research fields, i.e. the long tail. Such infrastructures and support services can range from consultations for researchers to help writing data management plans to the development and operation of data preservation and publishing services. Ideally, data infrastructures allow research data to be stored, managed, annotated and curated in a digital repository available 24/7 and to be used by multiple disciplines.
This paper presents the RADAR - Research Data Repository - project. With RADAR, a generic research data infrastructure for data preservation and publication in the above-mentioned fields of the long tail of science will be developed and established.

2 RADAR - Research Data Repository

RADAR is a generic end-point repository for digital research data providing both, preservation and publication services, thus allowing re-use of data. Data from various long tail disciplines can be stored in RADAR, i.e. natural, engineering and cultural sciences as well as humanities. The repository is developed in the course of a project funded by the German Research Foundation (DFG) from 2013 to 2016.
Further information is also available on the website: http://www.radar-projekt.org

2.1 Project partners

RADAR is developed within a national cooperation of research institutes from the fields of natural and information sciences: The technical RADAR infrastructure is provided by FIZ Karlsruhe – Leibniz Institute for Information Infrastructure and the Steinbuch Centre for Computing (SCC), Karlsruhe Institute of Technology (KIT). The Ludwig-Maximilians-Universität Munich (LMU), Faculty for Chemistry and Pharmacy, and the Leibniz Institute of Plant Biochemistry (IPB) provide the scientific knowledge and specifications and ensure that RADAR services can be implemented in the actual scientific workflows of academic institutions and universities. The sustainable management and publication of research data with Digital Object Identifier (DOI) assignment is provided by the German National Library of Science and Technology (TIB).

2.2 Goals of RADAR

The project partners aim to establish an interdisciplinary research data repository, which is sustained by research communities and based on a stable business model. In a nutshell, RADAR provides

- guidelines for researchers to introduce and facilitate research data management in general and to store/publish data in particular,
- secure data preservation in compliance with required storage periods (including permanent storage) using distributed data storage mechanisms,
- (optional) data publication with DOI assignment to secure traceability, access for re-use and citability, and
- technical implementation support for institutions (e.g. front end branding, data peer-review options for journals).

2.3 RADAR architecture

The RADAR 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 centers 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: The back end regulates general tasks such as access control or storage access, whereas the front end manages RADAR-specific workflows. These workflows include various data services: bitstream preservation, regular reports on data integrity, access control, data ingest processes, as well as the licensing for re-use and publishing of research data with DOI.

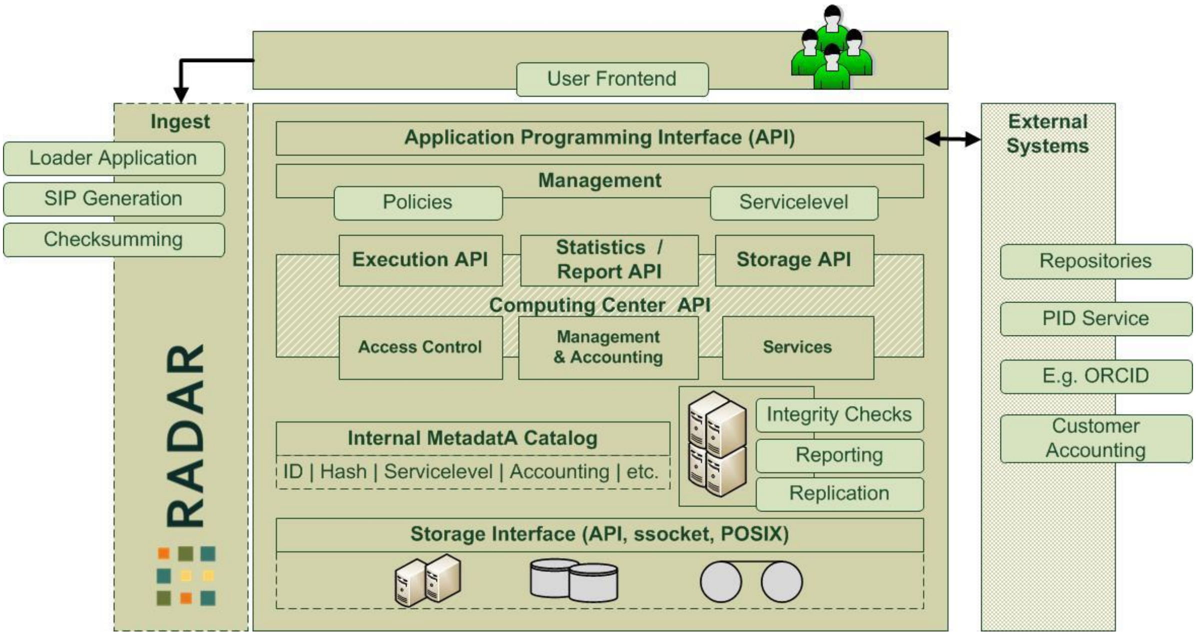

Fig. 1: Scheme of the RADAR architecture with data ingest and the API structure.

2.4 RADAR services - a closer look

The submission and integration of scientific data into repository collections is a continuing challenge for researchers and their affiliated research institutions. To facilitate this process RADAR offers detailed author guidelines and step-by-step explanations on how to choose between the offered preservation and publication services, how to prepare and how to submit the data. Depositing research data in RADAR ensures that the requirements of funding agencies and of Good Scientific Practice are met. Thus, other researchers will be able to find, re-use and cite published data. Published data will be assigned a persistent identifier (DOI) which increases their citability and can be listed in the researcher's citation record. As a generic, interdisciplinary service RADAR will accept all types of digital data that are collected in the course of scientific research studies. A dataset deposited in RADAR may comprise raw data, primary data (intermediate working data), secondary data and files describing the data and documenting the research process. Notably, RADAR accepts both data underlying scientific articles and standalone data publications, e.g. 'negative data'. RADAR does not accept pre-prints, doctoral theses or other grey literature. However, if it is part of raw data used for analyses, RADAR strongly encourages data depositors to provide information on related content using the metadata schema (see Tab. 1, parameter 'related identifier'). Data may be submitted in any file format to the service packages described in the following paragraphs (2.4.1 & 2.4.2). However, format recommendations will be provided in the author guidelines to facilitate data re-use; e.g. the use of PDF/A or XML based formats for text files.
In this context, the heterogeneity of research data is a serious issue for many research data repositories, especially when they provide storage and publication services for a wide range of scientific disciplines. RADAR is facing this problem by focusing on real scientific workflows and elaborates a generic best practice approach that will be evaluated and tested with data provided by scientific partners from different research areas. RADAR pursues a two-stage approach with a discipline-agnostic basic service for preserving research data (2.4.1) and an extended service for data publication (2.4.2).

2.4.1 Service: Data preservation

RADAR offers format-independent data preservation with a set of minimum metadata. Data providers are given the opportunity to store their data in compliance with specified long-term storage periods (e.g. 10 years, according to DFG recommendations). This includes a safe preservation of up to 15 years with the data remaining unpublished. By default, the associated metadata will not be published, unless specified otherwise by the data provider. In addition, a flexible data and metadata access management will be offered, 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. The bitstream preservation offered by RADAR will produce safety copies of the resource to assure its preservation.

2.4.2 Service: Data publication

For making data citable and re-usable, RADAR offers a combined service of research data publication and permanent preservation. Datasets published in RADAR receive a Digital Object Identifier (DOI). With this persistent identifier, datasets can persistently and unambiguously be referenced. The service also includes an optional embargo period for the publication (=download availability) of submitted data that can be subsequently prolonged if necessary. The metadata describing the dataset will be published already during the embargo and datasets will be allocated a DOI. This ensures that datasets can be found and cited already when they are deposited, while downloads will only be possible when the embargo has expired.

This two-staged service structure offers an enhanced data management service for data providers (Fig. 2). A technical quality control of research data and corresponding metadata is performed during the transfer of uploaded data objects into RADAR. Data providers will be notified through both their e-mail and user account when the preservation and, if applicable, the publication process was successfully completed. The consistency of datasets preserved in RADAR is regularly be checked and documented. To assist researchers in the submission of descriptive metadata, the repository service will include detailed author guidelines along with appropriate examples from various research disciplines. Datasets can be retrieved by the data providers at any time after deposition.

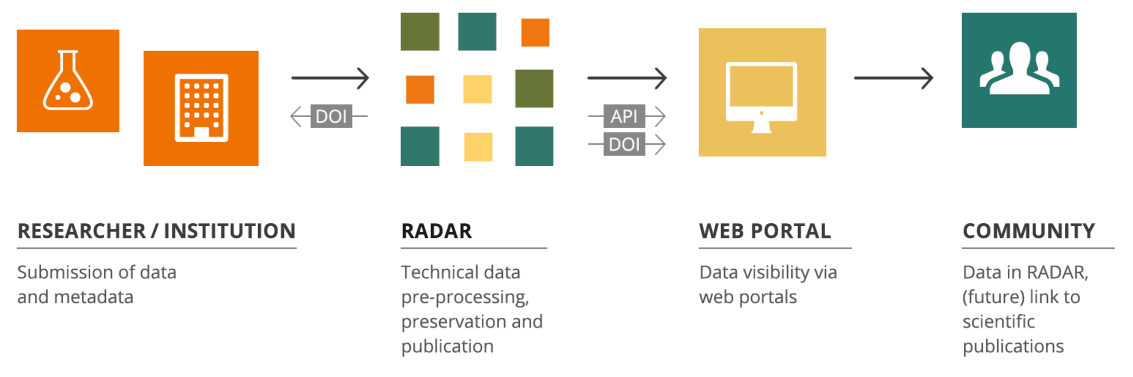
DOI
API
DOI
RESEARCHER / INSTITUTION
Submission of data and metadata
RADAR
Technical data pre-processing, preservation and publication
WEB PORTAL
Data visibility via web portals
COMMUNITY
Data in RADAR, (future) link to scientific publications

Fig. 2: General scheme of the RADAR service model. The goal: providing a sustainable preservation & publication service for research data.

3 RADAR: Data management for institutions & researchers

New mandates of several funding agencies require a formal data management. Such mandates have been implemented e.g. by the National Science Foundation (NSF) in January 2011 and require to include a data management plan within funding proposals (NSF 2011). In Germany, comparable developments are increasingly expressed in the form of scientific recommendations, e.g. in the guidelines for ‘Safeguarding Good Scientific Practice’ of the German Research Foundation (DFG 2013) and the decisions of scientific boards such as the Wissenschaftsrat (‘Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020’, WR 2012). The challenges that result from these mandates include the preservation and publication of research outputs so they can be re-used by fellow scientists (and the public) anywhere and at anytime.

The interdisciplinary data repository RADAR meets this challenge at various stages: While archived datasets are only available for the data providers, if not otherwise specified, all published data is open access and available via the corresponding landing pages. Published data will receive a DOI. Users can download data from RADAR as soon as the datasets are published, or, if applicable, as soon as the publication embargo has expired. This allows fellow scientists to test and reproduce the findings, to re-use and combine the data. Consequently, new hypotheses can be imagined and tested. As part of the data publication service, data providers may receive regular reports on usage statistics such as the number of downloads. The terms on which data may be re-used depend on copyright laws in effect and appropriate licences which have been assigned to data upon deposition in the RADAR repository. For published datasets, the usage of standardized Creative Commons (CC) licences (version 4.0) is recommended. However, customized licences or other descriptions that specify the terms of re-use may also be given. The licence information will be displayed to users on the landing pages, together with other descriptive metadata such as author(s), title, year of publication, DOI, related information and the download link.

3.1 How can institutions use RADAR?

- Depending on individual requirements, clients and institutions planning to use RADAR for data preservation and publication will be provided an account along with a storage contingent and administrator rights to access the services.
- Institutions can authorize their staff to use RADAR e.g. as data curators; they may choose to archive and optionally publish their data.
- Institutions may ask scientists about their data: Where is it? How are they planning to preserve it? Researchers must consider that underlying data is important. This refers to data storage and publication beyond the boundaries of research labs and specific projects. Institutions should highlight that this data is also part of the scholarly record and deserves the same care and attention as their articles. In this context research teams and established academics can also use RADAR to archive data also from past years of published or unpublished research (‘old data’ and ‘negative data’).
- Institutions may highlight that scientists can gain additional visibility from making their data files available and citable via a repository. Furthermore, scientists are encouraged to add a section for data to their CV’s.
- Institutions may refer authors of grant applications to RADAR when they are drafting their data management plan. Researchers may identify RADAR as the destination repository for archived and published data from their research.

3.2 How does data get into RADAR?

- After registration, an RADAR account along with a storage contingent and administrator rights to use the RADAR services is created.
- Within the RADAR user interface, roles for data management can be assigned. This includes e.g. roles for administrators and data curators. Institutions may provide graded access rights to processed datasets, as well as to the corresponding archived or published data.
- Following the log-in, so-called ‘workspaces’ for data management can be created. These workspaces can be assigned to specific persons (=data curators) who are responsible for data upload, the provision of background information (metadata) as well as data structure and form.
- After the dataset is completed in the RADAR workspace, the responsible curator may chose between the data preservation (5, 10 or 15 years storage duration) or publication.

3.3 Future services

With the start of the third project year in summer 2015, the implementation of additional services is planned. These include a bi-lingual user interface (English and German) and data peer-review services for journal editors and reviewers. Furthermore, RADAR anticipates merging its data submission process with manuscript submission systems of co-operating publishers. While uploading scientific manuscripts, authors may use RADAR at the same time for data provision to editors and reviewers. After acceptance of the manuscript, a two-way DOI-based linking between the article and its respective data increases visibility and citability of both, scientific paper and research data. Moreover, technical support services for institutions to provide e.g. front end branding are planned.

4 Business and cost model

The business model ensures a sustainable operation environment for the data archive as well as a tool for scientists to apply for data management funds. This model includes one-off payments by data providers, depending on data volumes and storage periods. This assures the projects financial independence from third-party companies or institutions and consequently its long-term sustainability. A cost calculation tool will enable researchers and institutions to receive cost estimates before using any of the offered storage services. With this tool, RADAR provides the opportunity to analyze costs for data preservation and publication during the project planning phase and to implement these estimates in data management plans. Furthermore, researchers are encouraged to include the estimates of costs in grant applications to receive funding for research data publication and the associated preservation services. With this approach, RADAR follows the increasing requirement of research outcome to be open access.

5 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 parameters 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 (Tab. 1) includes nine mandatory fields which represent the general core of the scheme. These fields contain the main requirements for DOI registration, in accordance with the DataCite Metadata Schema 3.1 (DataCite 2014) and must be supplied when submitting metadata to RADAR. Additionally, 12 optional metadata parameters serve the purpose of describing discipline-specific data. The parameters were implemented with a combination of controlled vocabularies and free-text entries, thereby covering heterogeneous data produced by a multitude of disciplines. The controlled vocabulary entries were defined in accordance with established regulations in mind (for example, ISO standards for language and country of origin of the data). 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 (CreativeCommons.org 2014). RADAR will actively disseminate all published metadata to DataCite. Metadata of datasets that are only stored (not published) in RADAR (Data Preservation, 2.4.1) 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.

Descriptive metadata: nine standard properties	**Mandatory parameters for basic information**
identifier	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	Year, in which data was created or refers to
subject area	Scientific fields appropriate for the resource
resource	Resource's content (e.g. dataset, model, software)
rights	Rights management statement (e.g. CC BY)
rightsholder	Institution/Person holding rights
Descriptive metadata: twelve optional properties	**Parameters for discipline-specific data descriptions**
additional title	Complementary textual information

description	Further information (e.g. abstract)
keyword	Keywords describing the subject focus
contributor	Associated institution/person (e.g. funder)
language	Main language used or relevant to resource
alternate identifier	Unique string within its domain of issue (e.g. local identifier)
related identifier	Identifiers of related resources
geolocation	Region/Place where resource originated/refers to
data source	Data origin (e.g. instrument, observation, trial)
software application	Software used for data production and processing
data processing	Specifies further processing (e.g. statistics)
related information	Further information (e.g. database number)

Tab. 1: Descriptive RADAR Metadata Schema with a set of generic parameters to allow an accurate and consistent identification of a resource for citation and retrieval purposes. The optional parameters can be applied accordingly, in order to meet the requirements of discipline-specific datasets.

6 Summary and Conclusions

As funding agencies adopt new mandates for sustainable research and access to underlying data, scientists are challenged to participate in a more active research data management and seek institutional support. In response, academic libraries start to extend their professional record of managing knowledge resources towards the area of research data and offer data management skills and infrastructure resources. Since 2013, the German National Library of Science and Technology (TIB) is a partner in the DFG-funded three-year collaboration project RADAR. The project develops a solid infrastructure for managing ‘long tail’ research data that do not fall within the scope of big data or discipline-based repositories. Offering data preservation and publication services, RADAR is an interdisciplinary digital data repository to be used by research facilities and libraries as a platform for data management. In a two-staged service model, clients choose if they either want to securely store data (non-public) or publish data as part of their research documentation. Both can be done in the RADAR workspace, requiring only a single login/authorization from their institution. RADAR services include DOI assignment for published data, a discipline-agnostic metadata schema and estimates of cost for grant applications, amongst others. The services will support researchers and their corresponding institutions in active data management: Data deposition in an appropriate repository serves as the basis to allow data re-use and to fulfill the research data life cycle. Moreover, appropriate tools to structure, store, catalogue and publish data may open new doors and innovative approaches of ‘doing science’: New co-operations can be formed and re-formed within and across disciplines and countries. The benefits may include added credit and value to research data, along with new insights, observations and policy preferences.

Acknowledgements

We wish to thank the RADAR project team for their input to this manuscript and gratefully acknowledge the tremendous efforts of all people providing support to the project up front and in the background. This project is funded by the German Research Foundation (DFG, Project No.: BE 1042/7-1, HE 5985/4-1, NE 1352/2-1, RO 2273/4-1, WE 1467/14-1).

References

CreativeCommons.org (2014). CC0 1.0 Universal (CC0 1.0) Public Domain Dedication. Retrieved from http://creativecommons.org/publicdomain/zero/1.0/deed.en. 2015-05-05.

DataCite (2014). DataCite Metadata Schema for the Publication and Citation of Research Data. Version 3.1 doi:10.5438/0010

Deutsche Forschungsgemeinschaft DFG (2013). Safeguarding Good Scientific Practice. Recommendations of the Commission on Professional Self Regulation in Science, Wiley-VCH.

European Commission (2013). Research Data e-Infrastructures: Framework for Action in H2020. Retrieved from http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/framework-for-action-in-h2020_en.pdf. 2015-05-05.

Harris, S.J. (2011). Long-distance corporations, big sciences and the geography of knowledge. In S. Harding (Ed.), *The Postcolonial Science and Technology Studies Reader* (pp. 61-83).

Heidorn, P.B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, *57*(2), 280-299. doi: 10.1353/lib.0.0036

Klump, J. (2009). Managing the Data Continuum. Retrieved from http://oa.helmholtz.de/bewusstsein-schaerfen/workshops/kolloquium-the-data-continuum-managing-sharing-and-re-use.html. 2015-05-05.

National Science Foundation NSF (2011). Proposal Preparation Instructions, Chapter II. In *Grant Proposal Guide.* Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp. 2014-10-17.

Neuroth, H., Strathmann, S., Oßwald, A., Scheffel, R., Klump, J., Ludwig, J. (Eds.) (2012). Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme. URN: http://nbn-resolving.de/urn:nbn:de:0008-2012031401

Science Editorial (2011). Challenges and opportunities. *Science, 331*, 692–693. doi:10.1126/science.331.6018.692

Thessen, A.E., Patterson, D.J. (2011). Data issues in the life sciences. *Zookeys, 150,* 15-51, doi:10.3897/zookeys.150.1766

Treloar, A., Harboe-Ree, C. (2008). Data management and the curation continuum. How the Monash experience is informing repository relationships. In *14th Victorian Association for Library Automation*. Retrieved from http://arrow.monash.edu.au/hdl/1959.1/43940. 2015-05-05.

Wissenschaftsrat WR (2014). Empfehlungen des Wissenschaftsrats zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020. Retrieved from http://www.wissenschaftsrat.de/download/archiv/2359-12.pdf. 2015-05-05.

Whyte J, Stasis A, Lindkvist C (2015). Managing change in the delivery of complex projects: Configuration management, asset information and ‘big data’. *International Journal of Project Management*, doi:10.1016/j.ijproman.2015.02.006