Project Proposal to Deutsche Forschungsgemeinschaft
in the Programme
Scientific Library Services and Information Systems

Funding Programme
Information Infrastructures for Research Data

Continuous quality control for research data to ensure reproducibility: an institutional approach

Acronym: "CONQUAIRE"

Prof. Dr. Philipp Cimiano
Semantic Computing Group
Center of Excellence Cognitive Interaction
Technology (CITEC)
Universität Bielefeld
cimiano@cit-ec.uni-bielefeld.de

Prof. Dr.-Ing. Gerhard Sagerer
Rektor
Universität Bielefeld
rektor@uni-bielefeld.de

Summary

Reproducibility is a cornerstone of the scientific process. While the reproduction of an experiment can be extremely difficult, the ability to reproduce the (computational) analysis of the data that supported a certain conclusion (e.g. the validation of a hypothesis) should be a minimum requirement on every piece of published research. We call this type of reproducibility *analytical reproducibility*.

The ability to reproduce the analytic results of a certain piece of research requires, as a minimum, that: i) the primary or secondary data is available, ii) the data is syntactically well-formed and ready-to-use, iii) the data is appropriately documented, iv) the analysis procedures (e.g. scripts) that were used to process or analyze the data are available, and v) these analytic procedures can be run on the data to reproduce the actual result published in a paper. Analytical reproducibility is often hampered by the fact that one of the above requirements is not met.

The goal of this project is to extend the infrastructure available at Bielefeld University for the management of data and publications by a framework that supports researchers in meeting the above mentioned requirements and thus to make their work analytically reproducible. Departing from current practices where data and software is published at the end of a research project, if at all, we intend to move the hosting of data to the very beginning of the scientific process. Borrowing ideas from computer science and from continuous integration, we intend to implement a continuous quality control framework that from early on encourages researchers to publish their data and analytic procedures so that these can easily be re-used and verified. The way we understand quality of data in the project is thus in the sense of *readiness to be re-used* and validated.

Towards this goal, we will interact with a selected group of researchers at Bielefeld University that have committed themselves to define a use case, provide requirements, implement pilots, and continuously work with the infrastructure and provide regular feedback. The researchers come from disciplines as varied as psychology, sports sciences, biology, chemistry, cognitive linguistics, computational linguistics, robotics as well as economics. By involving a varied set of disciplines, our goal is to identify common requirements on an infrastructure that supports data quality as a continuous process, and supports sharing and external validation of research results.

Besides extending our infrastructure, the project can be expected to have an impact way beyond Bielefeld University. By sharing our experiences and requirements identified, we hope to inform other universities and policy makers on the trade-off between effort and return-on-investment and which policies to adopt to support higher transparency in research.

Contents

1 Starting point and preliminary work

Reproducibility of scientific results is a cornerstone of science. Karl Popper wrote [16, p. 45]:

> *"We do not take even our own observations quite seriously, or accept them as scientific observations, until we have repeated and tested them. Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated coincidence, but with events which, on account of their regularity and reproducibility, are in principle intersubjectively testable."*

This ability for constant, rigorous testing and validation of research results ensures the integrity and efficiency of science. Stating it in the words of the OECD [14]:

> *"Sharing and open access to publicly funded research data not only helps to maximise the research potential of new digital technologies and networks, but provides greater returns from the public investment in research."*

An illustrative example of data sharing and independent validation or refutation of scientific results by others is conveyed by a recent case: In 2010 Harvard researchers Kenneth Rogoff and Carmen Reinhart published their paper "Growth in a Time of Debt" [18]. They had analyzed historical data from 20 industrialized countries since WWII and concluded that economic crises arise when the size of a country's debt rises above 90 % of the Gross Domestic Product (GDP). This result had a significant impact on political decisions worldwide, until in 2013 the student Thomas Herndon tried to replicate the analysis. He contacted Reinhart and Rogoff and they provided him with the actual working spreadsheet. Inspection of this spreadsheet disclosed several serious errors, which rendered the results invalid [6].

The above case is thus a very good example of the efficiency of the scientific process, albeit a rather simple one in which both the data and the analytical procedures used to analyze the data were available in the form of a spreadsheet and could thus be reproduced in a straightforward fashion. In many scientific disciplines, reproducing research results is more complicated as research can involve advanced and high-tech devices needed to measure certain phenomena, complex experimental protocols, data in different formats (structured vs. unstructured), different modalities (text, annotations, video, audio, 3D data) etc. which make it difficult to reproduce a certain scientific experiment.

While reproducing an experiment can be very challenging, as a baseline, given the primary or derived data resulting from an experiment, the reproduction of the (computational) analysis procedures that yielded a particular result should be feasible. In fact, an important step in the generation of scientific results lies in the computational analysis of the primary data or derived secondary data. In most cases, software packages (such as SPSS, R, Excel) are used in this part of the process to test a hypothesis by performing some computational or statistical analyses of the primary or derived data. We will refer to this part of the process as *analytical phase*. While being able to fully reproduce an experiment can be extremely difficult, reproducing the analytical phase seems more feasible as it would essentially require access to the primary and/or derived data as well as to the analytical tools used by the researchers to derive some result.

Thus, a significant step towards supporting reproducibility in science would be "analytical reproducibility", which consists of making sure that a third party researcher could reproduce the computational and statistical analysis performed on primary and derived data to yield a particular conclusion, thus being able to independently verify the results and conclusion. A crucial question is how research data infrastructures should be extended to support analytical reproducibility, sharing and thus independent validation of (analytical) research results. As a prerequisite for a research result or scientific paper to be analytically reproducible and useful for other researchers, the following conditions need to be met:

- primary or secondary data is available for inspection and processing,
- data is syntactically well-formed and follows best practices from the corresponding community, thus being *ready-to-use*,
- analytic procedures (e. g. scripts, spreadsheets, etc.) that were used to process or analyse the data are available, and
- these analytic procedures can be run on the data to reproduce the actual result published in a paper.

Analytical reproducibility is often hampered by the fact that one of the above four elements is lacking. In order to be reproducible, research data needs to have a certain *quality* which we operationalize in the context of this proposal as its *readiness* to be re-used by others, e. g. to reproduce the computational analyses described in a scientific publication. Typically, researchers have no institutional support nor resources to ensure data quality in the above sense. Thus, curating data to make it fit for publication and sharing requires substantial resources that researchers do not typically receive credit for. If the data is published, this is typically done in a delayed fashion after the research project or dissertation work has been concluded. However, it is well-known from other areas (e. g. software engineering) that quality control is better taken into account from the start of the project. Continuous integration [2, p. 209] in software engineering is aimed at increasing quality of software by specifying a number of tests that the software should pass in order to continuously monitor compliance with these. Drawing inspiration from software engineering, principles of continuous integration could be applied to research data management to realize continuous monitoring of data quality and ensure that at each step in the research cycle, the data fulfills a number of defined tests. Ultimately, as a final test on the quality of the research data, the proof that third parties can reproduce and validate the (computational) analyses that produced a certain result could be seen as final culmination of a continuous data quality assurance cycle.

In our previous efforts [20, 10, 5, 22] in research data management we learned that researchers are generally willing to create data of high quality, share this data and make their results reproducible as part of their duties as a researcher and to meet expectations of their community. However, the challenges related to data processing, validation and publishing are rather demanding. Therefore, researchers need to be supported in this process by an appropriate institutional infrastructure that hosts their data and implements corresponding workflows that allow to ensure research data quality and reproducibility along the whole research lifecycle. Such an infrastructure that supports continuous research data quality monitoring and at the end makes the data publicly available to allow reproduction of the computational analysis is not yet available, and this proposal aims to close this gap.

Related Work

Policy context

More and more journals, funders and research organisations encourage successful data and code sharing practices. A recent study has shown that a growing number of **international journals** require authors to provide sufficient detail to reproduce results described in submitted manuscripts [19]. Examples are the journals *Biostatistics* [15] and *PLOS ONE*[1], which enforce publication of research data including both the datasets and scientific software from which results were derived. Some journals provide guidance on how scientific software can be peer-reviewed and automatically tested [8].

Research funding agencies have also issued corresponding policies to enforce data disclosure. For instance, DFG's Rules of Good Scientific Practice include the directive to store data and materials that

[1] `http://www.plosone.org/static/policies`

have contributed to research findings for 10 years [4]. In its framework Horizon 2020, the European Commission requires Open Access to scientific publications as a general rule and to research data in particular fields.[2]

Journal and funder directives inform **data management policies at universities**. The League of European Research Universities (LERU) has suggested a Roadmap for research data noting the important linkage of policies addressing research data, technology and support [1]. A recent German Rectors' Conference (HRK) recommendation perceives the management of research data as key strategic challenge for university management.[3] Accordingly, a coordinated approach is needed aiming at institutional research data policies to be backed by support structures for data literacy and research data management. These support structures have to address data management planning, data publication and long-term preservation of data. According to the HRK recommendation, only Bielefeld University has passed such directives. Recently, the University of Göttingen, Humboldt University of Berlin and University of Heidelberg have followed.

Scientific data and software publication

To comply with these data disclosure policies, more and more researchers use open source software hosting facilities such as GitHub or SourceForge. These services are becoming increasingly important for managing various types of research outputs, ranging from datasets to statistical code and even complete manuscripts [17].

In order to assess the uptake of open source software hosting services used for disclosing data in scientific literature, we searched for mentions of these open source hosting services within the Europe PubMed Central (Europe PMC) full text subset. Europe PMC comprises access to more than 2.6 million life science and biomedical research articles. Figure 1 shows the quarterly distribution of five selected hosts. Data was collected on 28 August 2014. Absolute figures suggest that open source software hosting facilities are gaining in importance in general, and GitHub in particular.

To reflect this development, the general purpose Open Access repository Zenodo, created by the EC-funded project OpenAIRE and CERN, allows to archive software releases from GitHub and assigns a Digital Object Identifier (DOI). In Germany, the DFG-funded SciForge project explores publication and citation of scientific software with persistent identifiers. It has published ten open issues to make dynamic data and code development better citable through research information infrastructures.[4]

Even though promising steps have been taken, safeguarding computational experiments' details is still beyond the scope of today's research data infrastructures at German universities. Institutional repositories only preserve static copies of files.[5] However, successful data and code sharing in computational research requires that all analytical procedures are logged and available. It is also important to permit collaborative review to scientific software development including the management of changes to the code. Given its growing coverage, interoperation between institutional repositories and services such as GitHub or SourceForge are key to increase the visibility of computational research findings within the academic domain. After all, advocacy and incentives for data and code sharing need to be implemented

[2] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

[3] http://www.hrk.de/resolutions-publications/resolutions/resolution/convention/management-of-research-data-a-key-strategic-challenge-for-university-management/

[4] http://www.sciforge-project.org/2014/05/19/10-non-trivial-things-github-friends-can-do-for-science/

[5] Four German universities offer data publication through institutional research data repositories so far. Open Data LMU <http://data.ub.uni-muenchen.de/>, MADATA <https://madata.bib.uni-mannheim.de/>, PUB <http://pub.uni-bielefeld.de>, HeiDATA <https://heidata.uni-heidelberg.de/>

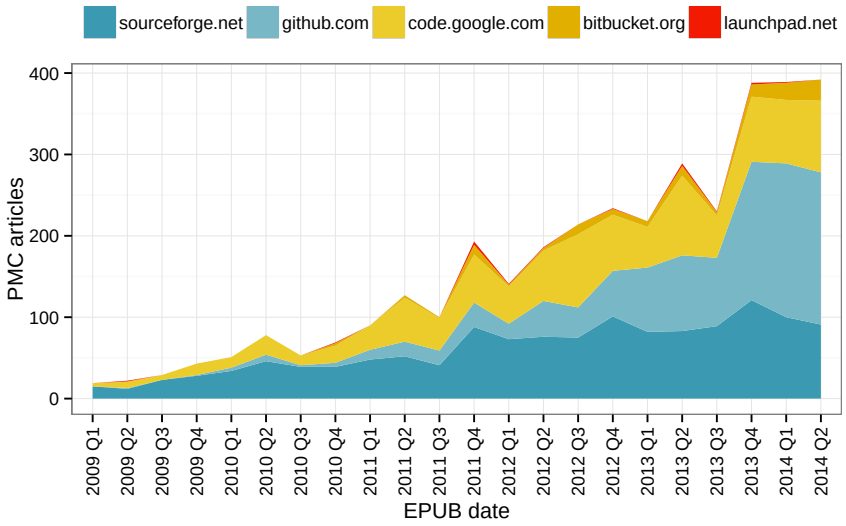

Figure 1: Quarterly distribution of links to open source hosting facilities in Europe PMC full text articles

across the institution.

Preliminary Work

Bielefeld Data Informium

In 2009, the Rektorat of Bielefeld University initiated the *Bielefeld Data Informium* project. It has the goal of providing an efficient and effective research data management infrastructure at Bielefeld University.

Informium's activities operate on a multidisciplinary scale and address research data management from two different angles (see Figure 2): Due to complex requirements and heterogeneous methods in the disciplines, the bottom up strategy incorporates researchers and embedded data managers from different disciplines. Novel collaborations between main research fields and information infrastructure facilities have been successfully established. Among them are the Data Service Center for Business and Organizational Data (DSZ-BO), the Collaborative Research Centre (SFB) 882 "From Heterogeneities to Inequalities", the Institute for Interdisciplinary Research on Conflict and Violence (IKG), and the Center of Excellence Cognitive Interaction Technology (CITEC). The top-down strategy consists in analyzing and inferring general requirements for the institutional research data infrastructure. The operative monitoring is done by the **focus group**, an advisory group consisting of researchers and other university stakeholders (e. g. the Vice-rector for research, young researchers and transfer, the CIO-IT, the Head Librarian) who counsel and oversee the activities.

As a result of Informium, Bielefeld University has passed guidelines and policies on research data management, being the first German university to form an institution-wide agreement upon standards for research data handling among all stakeholders.[6] The university-wide policy calls on researchers (i) to take advantage of the universityâĂŹs advisory services for research data management and (ii) to publish research data through registered research data repositories. To this end, Bielefeld University offers comprehensive advisory services on data management planning, and data publication and preservation through its institutional repository "PUB – Publikationen an der Universität Bielefeld".

[6] `https://data.uni-bielefeld.de/en/resolution`

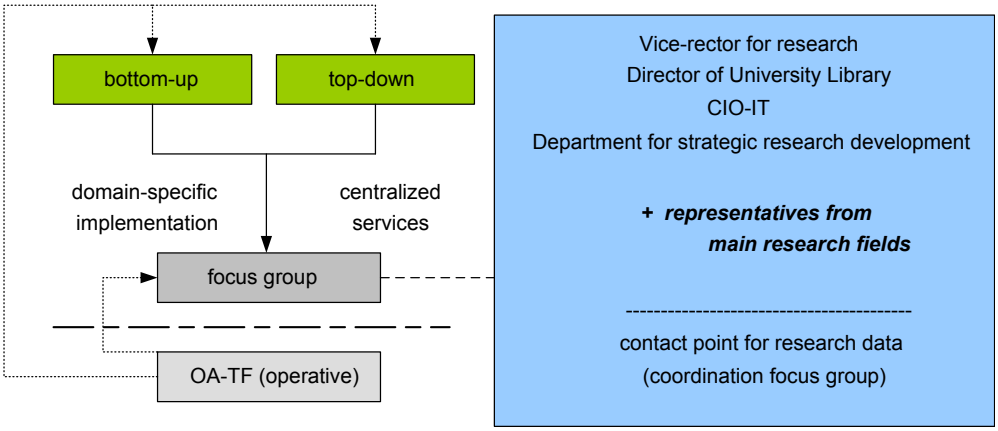

Figure 2: Informium strategy: bottom-up and top-down approach

Bielefeld University Library

According to the German university rankings and competitions (CHE, BIX) Bielefeld University library is one of the most innovative university libraries in Germany, leading both in terms of user satisfaction and on account of its pioneering role in the fields of Open Access, research data management, digital research services, and search engine technologies. Among ongoing projects and networking activities, prominent contributions to the international repository based research infrastructures are the Bielefeld Academic Search Engine (BASE) as well as the participation in EC-funded projects OpenAIREplus and EuropeanaCloud.

In a co-operative effort, Bielefeld University library is currently developing new types of research data infrastructures in the Collaborative Research Center (SFB) 882 “From Heterogeneities to Inequalities”. Bielefeld University Library has contributed to leading national and international initiatives by participating in guideline development for digital research repository networks (COAR, DRIVER/OpenAIRE, DINI), usage data federations (Knowledge Exchange), data metrics (RDA/WDS Data Publishing IG), Research Core Dataset for German and the Data Documentation Initiative (DDI). Bielefeld University Library contributes to international open source collaborations, including LibreCat/Catmandu and rOpenSci.

CITEC

The Center of Excellence Cognitive Interaction Technology (CITEC) is highly committed to the ideals of Open Science and endorses the view that the results of publicly funded research as well as the underlying research data are a public good that should be openly accessible to anyone. In 2013, CITEC’s Scientifc board passed the “Open Science Manifesto” [3], which describes CITEC’s strategy towards this goal and describes measures that will support the sharing of research data. CITEC employs its own research data manager (Cord Wiljes, involved in this proposal), who supports researchers in managing and releasing their research data. Since then, CITEC has consequently invested in building up an infrastructure that provides open access to research data and software. Further, CITEC also carries out active research on issues related to research data management and data quality [11, 12, 21, 22, 13].

1.1 Project-related publications

1.1.1 Articles published by outlets with scientific quality assurance, book publications, and works accepted for publication but not yet published

[FMzVP+13] Stefan Friedhoff, Christian Meier zu Verl, Christian Pietsch, Christian Meyer, Johanna Vompras, and Stefan Liebig. Replicability and comprehensibility of social research and its technical implementation, 2013.

[HLS+13] Maarten Hoogerwerf, Mathias Lösch, Jochen Schirrwagen, Sarah Callaghan, Paolo Manghi, Katerina Iatropoulou, Dimitra Keramida, and Najla Rettberg. Linking data and publications: Towards a cross-disciplinary approach. *International Journal of Digital Curation*, 8(1):244–254, 2013.

[KLS+12] Stefan Kramer, Amber Leahey, Humphrey Southall, Johanna Vompras, and Joachim Wackerow. Using rdf to describe and link social science data to related resources on the web, 2012.

[MWC14] John McCrae, Cord Wiljes, and Philipp Cimiano. Towards assured data quality and validation by data certification. Proceedings of the LDQ 2014 Workshop, 2014.

[WC12] Cord Wiljes and Philipp Cimiano. Linked data for the natural sciences: Two use cases in chemistry and biology. Proceedings of the Workshop on the Semantic Publishing (SePublica 2012), pages 48–59, 2012.

[WJL+13] Cord Wiljes, Najko Jahn, Florian Lier, Thilo Paul-Stueve, Johanna Vompras, Christian Pietsch, and Philipp Cimiano. Towards linked research data: An institutional approach. In Alexander García Castro, Christoph Lange, Phillip Lord, and Robert Stevens, editors, *3rd Workshop on Semantic Publishing (SePublica)*, pages 27–38, 2013.

2 Objectives and work programme

2.1 Anticipated total duration of the project

36 months starting from 1 August 2015.

2.2 Objectives

Analytical reproducibility of research results as described in the introduction, requires:

- **Availability of Data:** the data (primary or derived, depending on specific case) should be available in some common and open format, syntactically valid and properly documented so that independent researchers can make sense of the data and work with it using standard tooling.
- **Analytical Reproducibility:** the computational procedures used to process and analyze the data should be available in such a way that they can be executed, analyzed and validated.
- **Quality & Compliance:** agreed-upon quality criteria and constraints (e. g. in order to enforce that measurements are within a given range or annotations follow a given set of guidelines) should be made explicit so that compliance with respect to these can be verified.

A crucial aspect of the above mentioned points is thus monitoring data quality, adherence to best practices of the community, passing semantic integrity tests, etc. We will favour an agile approach to research data management in which data quality is continuously monitored and ensured right from the start of the research cycle. So far, in most scientific projects, data publication is typically considered only when the research is actually finished and results have been published in a written article. In general, there are few incentives for researchers to invest in data publication, curation and enhancement to make their analyses and results reproducible by others once the research project has concluded. Instead, we intend to move data publication to the heart of the research process. Our goal is to support researchers by an appropriate workflow and incentive creation mechanisms that support them in this task of making their analyses traceable and executable[7].

[7] See also `http://www.executablepapers.com/` for a closely related vision of an "executable paper", but also [9] and [7].

In order to maximise acceptance, we envision a user interface so unobtrusive that it is invisible most of the time. When a researcher generates a new project, a directory is created and populated with a few suggested files and sub-directories. From then on, this directory hierarchy is automatically synchronized (in the background) with a versioning file server that also performs validation tests and gives feedback when required (i. e. tests have failed).

At all stages, a researcher can view a visual representation of the status of the project towards achieving full analytical reproducibility, with indication of actions to be performed to achieve the status of full analytical reproducibility including an indication of which tests the data has passed. Full reproducibility will only be achieved if a third party validates that the results of the analytical procedures run on the data provided actually return the results as described in a given paper.

Beyond the technical support for hosting that we provide, there are clear incentives for research to take part in this. Generally, researchers are interested in data quality and making their results reproducible. By supporting them with validation services, flaws in the data can be spotted early on, thus contributing to cleaner data and clearly documented analytical procedures. So the main incentive for researchers during the project will be that our services will support them in ensuring data quality and ensuring that the data follows certain constraints etc. from the very beginning. This will be supported by a gamification approach in which a dataset receives a number of badges when certain criteria are met (data availability in some open format, script availability, documentation, re-use of open and standard vocabularies, etc.). Defining this social reward system will be an important goal of the project.

Finally, for a researcher, allowing early validation by others (supported through the version control system) will help to ensure that there are no hidden nor unwarranted assumptions nor flaws in data collection, analysis etc. This will contribute to higher quality research results that researchers strive for.

The goal of the project is to extend the infrastructure for research data management available at Bielefeld University (see section 1, sub-section “Preliminary work”) to support analytical reproducibility in the above sense. Ensuring data quality in terms of *readiness to be reproduced* will be our main goal. Instead of leaving quality assurance to the end of the research cycle, we borrow inspiration from the paradigm of continuous integration (i. e. fast release cycles accompanied by automated unit tests for quality control) used in software engineering to make data quality assurance a continuous process rather than an afterthought as in most research projects.

Taking inspiration from this, our goal is to embed continuous quality control for research data into the existing infrastructure at Bielefeld University. The mechanisms for ensuring quality control and analytical reproducibility will be layered on top of a distributed version control system (DVCS) that will be rolled out university-wide as part of this project. Besides providing data hosting and versioning of their data to researchers as a basic service, several extensions on top of this DVCS system will be implemented to support data quality control in our sense, in particular:

- **Reproducible paper workflow:** A pre-defined workflow will guide researchers from the start on to reach the status of analytical reproducibility by recording the status of the project in the background and providing clear advice on what next steps should be followed.
- **Certification of readiness to use and quality monitoring:** A quality monitoring and certification framework will run in the background and will analyze the data as uploaded to the DVCS system and perform some basic checks to analyze if data is in some open format, if it is syntactically correct, if common and standard vocabularies have been used, etc. Additional integrity constraints can be defined and implemented as required by researchers in a specific application domain to ensure the semantic validity of the data.

- **Public and collaborative validation:** As true reproducibility can only be verified by another researcher reproducing the results, we will use social reward mechanisms (including gamification), to encourage users to report reproduction of results.
- **User experience:** Data uploading will happen in the background, barely noticeable once configured. A user interface will allow easy access to the data for both the researcher and external users and display the public status consisting of its automatically measured readiness to use and social metrics, such as downloads, ratings and reported reproductions.

In extending the infrastructure for research data available at Bielefeld University, we will ensure a close interaction with researchers as potential end users. In fact, we have the commitment from 9 research groups from areas as diverse as psychology, sports sciences, biology, chemistry, cognitive linguistics, computational linguistics, robotics as well as economics to participate in this project by contributing to the definition of use cases, requirements on the infrastructure as well as continuous usage of the infrastructure in their own work (see letters of support). We will organize workshops in which use cases and requirements on the infrastructure will be defined together with these researchers. The infrastructure will support the daily work of the researchers. We will gather their feedback at regular intervals. We seek to implement a pilot for each of these research groups and realize tight and agile implementation cycles in which the infrastructure continuously evolves to meet the needs of the involved researchers. At the end of the project, extrapolating from the single pilots, we will seek to compile best practices that are regarded as feasible by researchers on the basis of the experiences with the pilots that can be elevated to policies and best practices at the university level.

Some of the open research questions to be addressed in this project are:

- How can the concept of analytical reproducibility be supported from an infrastructural point of view?
- What is the associated cost and return on investment?
- Can principles from continuous integration be applied to management of scientific data to increase data quality and support its readiness to be used?
- How can researchers be motivated and incentivized by practical workflows in such a way that the support is not found obtrusive?

The contributions of the project can be classified into: technical, conceptual and policy-related:

- **Technical:**
  - **Data hosting and versioning:** Implement and roll out across the whole university a DVCS-based infrastructure for data deposition, hosting and versioning
  - **Graphical user interfaces:** Building on existing tools, develop a web interface that supports researchers in accessing the DVCS system and results of the validation service
  - **Workflow:** Develop a lightweight and non-intrusive workflow system that guides researchers in what the next steps could be, providing recommendations towards achieving “full analytical reproducibility”
  - **Data Quality:** Develop a framework for monitoring reproducibility and readiness of use of the data and analytical procedures and semantic integrity and compliance with best practices of data by running automated tests in the background
  - **Quality Checks:** Implement domain-specific and domain-independent quality checks that ensure compliance with agreed-upon quality criteria and community-specific and general best practices in data publication
  - **Social Reward**: Implement a gamification system that provides social rewards to re-

searchers for publishing reproducible research and for reproducing the research of others.

- **Conceptual:**
  - **Requirements:** Gather domain-specific and domain-independent requirements on the infrastructure to support analytical reproducibility
  - **Proof of concept:** Provide with our pilots the proof of concept that the goal of analytical reproducibility is feasible and can be supported by an appropriate technical infrastructure
  - **Feasibility:** Determine the trade-off between cost and return on investment for achieving analytical reproducibility, and verify feasibility on the basis of this trade-off
  - **Catalogue of quality criteria:** Develop a catalog containing domain-specific and domain-independent quality criteria that need to be considered in striving for analytical reproducibility
  - **Feasibility of social reward and reputation system:** Explore together with researchers the feasibility of a social reward and reputation system that provides credit to researchers for achieving analytical reproducibility, develop requirements on such a social reward system together with researchers.
- **Policy-related:**
  - **Better understand** today's requirements towards analytical reproducibility in computational research fields
  - **Liaise** with a wide range of researchers and infrastructure facilities across institutions
  - **Provide advocacy and incentives** for data and code sharing
  - **Develop and adjust institutional data policies** which define roles and responsibilities in the research data management process
  - **Generate recommendations and best practices** to support analytical reproducibility based on the experiences gathered in the pilots.

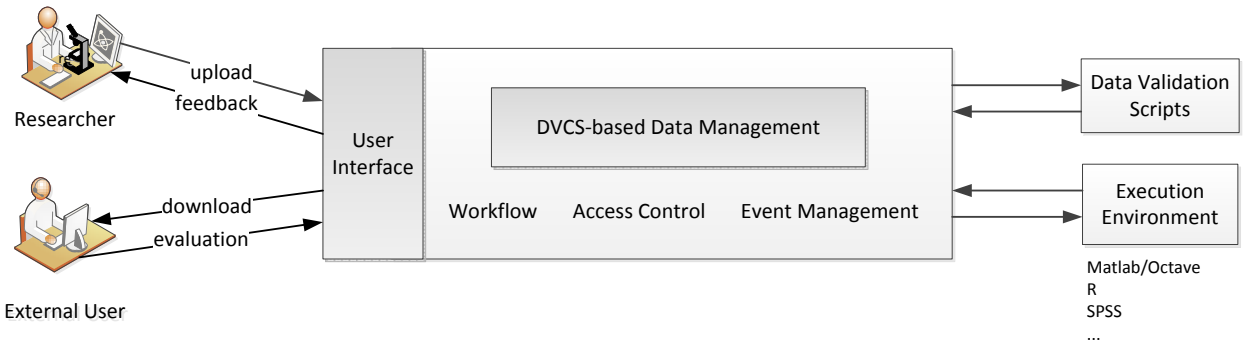

Figure 3: System Architecture

2.3 Work programme and proposed research methods

WP1: Case Studies

CONQUAIRE will be informed by partner projects from a wide range of disciplines, which will act as case studies to define initial requirements and enable a lean development with constant testing and user feedback. To prepare this proposal, we have already conducted an open workshop and subsequent individual conversations, in which and we acquired nine partner projects from a wide range of disciplines. These partner projects will ensure that the new infrastructure meets researchers' require-

ments in everyday work. Key goal of the infrastructure developed by CONQUAIRE will be usability and usefulness. Therefore, researchers from each partner project will be closely integrated during the whole project. Each of the nine partners was carefully selected for

- their interest and commitment to improve transparency in science and access to research data in particular
- their project's need to share data over institutional and/or disciplinary boundaries
- their willingness to publish their research data under an open license

The project partners have formally committed themselves to contribute to CONQUAIRE during its whole project lifetime, and in particular to:

- appoint a contact person who will be responsible for communication and planning
- participate in interviews and workshops to identify system requirements
- contribute to the definition of best practices
- provide research data
- continuously test the software and services developed by CONQUAIRE
- publish their research data at the end of their project under an open license, if possible

We consider the case study partners representative for research because they cover a wide range of disciplines. In addition, they differ with respect to the degree to which data management solutions are already in place. While most do not yet have their own solutions, some already implemented work-specific solutions. We intend to learn from these and will implement their functionality into CONQUAIRE as far as possible.

Based on the partner projects, we will define best practices for documentation and design appropriate forms that should be filled in by researchers to fully document their results and procedures. As part of this research, we will gather guidelines for documentation provided by external agencies in order to compile discipline-specific requirements on documentation, indicating which information is mandatory and which one is optional.

We will accompany each of the nine partner projects over a period of 24 months. The actual start will differ across projects. A timeline depicting the case studies is shown in Figure 4.

For each partner project a student assistant will be recruited, who will support the case study partners in their contributions to CONQUAIRE.

Case Study Partners

The following researchers formally agreed to participate with current research projects which will act as case studies to CONQUAIRE:

1. **Chemistry (Atmospheric Ice Nucleation):** In the context of the DFG research unit INUIT, the group of Prof. Dr. Thomas Koop carries out ice nucleation experiments with the same type of materials using diverse experimental techniques. CONQUAIRE will support data sharing between all involved groups to evaluate whether these data sets are internally consistent with each other, and finally to combine them, thus providing a parametrization that represents all data from the different techniques.
2. **Computer Science (Cognitive Service Robotic Apartment)**: The Central Labs Facility (CLF) lead by PD Dr.-Ing. Sven Wachsmuth participates in a large project with the goal to develop new approaches for intelligent smart homes that integrate service robots in order to interact with and support its inhabitants. The project constitutes a co-operation between 12 research groups, including computer science, linguistics, psychology and robotics, and therefore calls for cross-disciplinary data

sharing solutions.

3. **Biology (Navigational skills of bumblebees):** The group of Prof. Dr. Martin Egelhaaf investigates the abilities of bumblebees to return to their nest. CONQUAIRE aims to support the sharing of experimental data and software tools that will be shared with two international groups in order to test differing hypotheses about the bees' gaze strategy. By testing this novel form of international co-operation between otherwise competing research groups, data quality, processing and interpretation shall be improved and divergent theories shall be resolved. CONQUAIRE will thus contribute to higher transparency by enabling the comparison of results and hypotheses across two groups, making their analytics procedures explicit and thus directly comparable.
4. **Biology (Stick insect locomotion):** The group of Prof. Dr. Volker Dürr investigates the locomotion of different insect species using high-precision motion capture methods. CONQUAIRE will support Prof. Dürr in creating a database of locomotion data and to share this database with other researchers, and ensure the quality of the data.
5. **Linguistics (Natural language intermission):** Prof. Dr. Schlangen studies human/human verbal and non-verbal interaction in cooperation with the university Sorbonne in Paris. CONQUAIRE will support data sharing between partners and ensure the compliance with certain quality criteria and documentation best practices of the data.
6. **Economics (Agent-based economic simulation of a macro-economy):** The group of Prof. Dr. Sander van Hoog investigates macro-economic phenomena through agent-based simulations of economic environments consisting of many players. CONQUAIRE will support the sharing of data across associated research groups in Italy, UK, France, Turkey, Germany, and check compliance of the data with a set of pre-defined quality criteria.
7. **Sports Science (Mobile assistance system):** Prof. Dr. Thomas Schack pursues the goal of developing a mobile cognitive assistance system in form of an AR eye tracking glass that helps to lead a self-determined and independent life. This is a joint project between several scientific groups at Bielefeld University, the von Bodelschwinghsche Stiftungen Bethel, and several industrial partners. CONQUAIRE will support the sharing of the data with researchers and companies, and support the continuous quality control of the data.
8. **Psychology (Eye coordination in speed stacking games):** The group of Prof. Dr. Werner Schneider investigates the interaction between long-term memory and attention in speed stacking. It records multiple types of multimodal data, ranging from film and eye tracking data to event data and manual annotations. CONQUAIRE will support alignment and quality control of the data.
9. **Linguistics (Child language acquisition):** PD Dr. Katharina Rohlfing studies language acquisition of young children by investigating mother-child interaction during a game and a free play situation. Recorded data includes audio and video recordings from 3 camera perspectives, which will be extensively coded and transcribed. CONQUAIRE will support the sharing of the data with a partner group from the University of Warsaw, Poland, who will act as external validator of the data.

T1.1 *Requirement specification* (M1–6): At the project start we will conduct a general kick-off meeting with all project partners in which all partners will be informed about our plans in more detail, and a basic mode of operation will be defined. Subsequently we will establish communication channels with each of the groups mentioned above. Based on open interview sessions, we will define requirements and integrity conditions to inform the development of the overall platform. Based on these, a common requirements specification document will be created, which will guide the development of the first prototype early on, after six months.

T1.2 *Data management plans* (M1): Each partner project will create a detailed data management plan at the start of the research project and update it continually.

T1.3 *Continuous integration* (M1–36): Development throughout the whole project will be incremental as users inform us of their experience and provide feedback as a basis to refine the system. To achieve this, we will conduct interviews with researchers bi-weekly throughout the whole project. We will document which functionality or aspects the users regard as helpful, and which they regarded as non-useful or even obstructive. We plan to deploy the prototypes very early on in the project at M6 and then incrementally refine the prototypes as new requirements are fed to the project. From this month on, the whole infrastructure will be accessible by the participating researchers. The workflow system will be available in a first version by M12 and the execution framework by M24. Tandem partners will be involved early on in the project and have access to the data as soon as the researchers involved agree to provide them access, latest by M24.

T1.4 *Workshops* (M12, M24): After the first year and after the second year, we will conduct a formal workshop to document the experience and feedback of each of the use cases. These experiences will be documented as a part of a "lessons learned" report.

T1.5 *Data publication* (M36): The final goal of this work package is to fully open access to the data and analytic procedures by the end of the project.

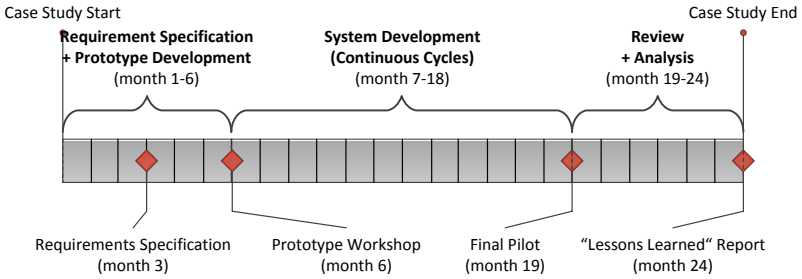

Figure 4: timeline depicting the case studies

WP2: Enhancing the Institutional Repository Ecosystem for Analytical Reproducibility

This work package examines how to track corresponding versions of data and code, and how to make them accessible and citable through institutional repositories. For this aim, we plan to fetch computational research artefacts stored in distributed version control systems (DVCS). They originate from web hosting facilities such as GitHub or are kept in self-hosted DVCS. The work package will furthermore align self-archiving of data and code with workflows for data validation and quality assessment (WP3, WP4). After all, this work package will not only assist our case studies (WP1). Rather, enhancing institutional repositories for computational research is a crucial step to encourage data and source code release across the university.

A timeline is shown in Figure 5.

T2.1 *Survey and deployment of DVCS* (M1–6): We will identify, compare and evaluate open-source software hosting facilities. The survey will incorporate requirements made by our case study partners (WP1). As a result of the survey, we will deploy one self-hosted solution that allows researchers to manage datasets and code stored in distributed version control systems (DVCS). It will be distributed to our case study partners and form the basis for the university-wide integration planning process (see 2.4).

T2.2 ***Design, implement and evaluate content deposition and ingestion strategy*** **(M4–24):** We will define tools and best practices for institutional repositories to interoperate with data and source code tracked through DVCS. The task will address both selected external web hosts and our self-deployed solution. Iterative interface design, implementation and evaluation will pay attention to single-sign-on, and vocabularies that a) describe activity streams, b) list code repositories, users and group accounts, and c) record source code changes. In the end, a suite of software modules to ease the import, storage and transformation of metadata about software repositories will be released through the open source community LibreCat[8].

T2.3 ***Adoption of data quality assessment framework*** **(M18–30):** The rating scheme that displays the levels of analytical reproducibility will be adopted (WP3).

T2.4 ***Interfacing the reproducibility workflow system*** **(M18–30):** Interfaces between the reproducibility workflow system (WP4) and our self-hosted DVCS will be maintained throughout the project.

T2.5 ***Dynamic citation*** **(M18–36):** To support the concept of analytical reproducibility, we will design and implement mechanisms to make corresponding versions of data and code citable through persistent identifiers. Persistent linkage will not only include each source code version, but needs also to be established to related literature, datasets and materials. In close collaboration with open repository networks (see WP5), shared identifiers for contributors and funders will be re-used. In particular, we will build upon DataCite Metadata Scheme, OpenAIRE Guidelines and standard linked open data vocabularies proposed in WP4.

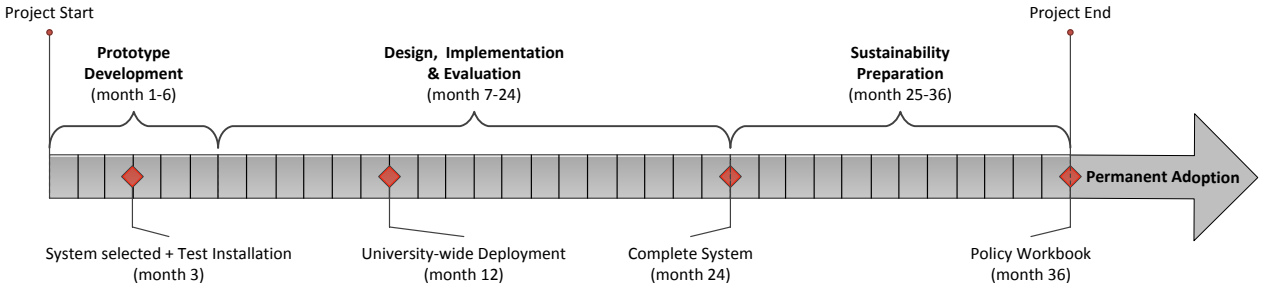

Figure 5: timeline WP2

WP3: Data Quality Assessment

This work package will implement a framework for data quality assessment. The goal will be to support data quality in terms of making data ready to be used by others. For this, a framework will be implemented in which domain-independent tests check syntactic wellformedness, validity of data according to schemas, consistent use of vocabulary, provision of metadata, and availability of scripts. Also, domain-specific quality checks will be implemented and integrated into the framework as services in a plug-in framework. The domain-specific quality checks will be defined together with researchers in the context of WP1 and implemented in the corresponding pilots. The developed framework will rely on continuous integration principles that help researchers to ensure from the start of the project that their data is *ready-to-be-used* and their analytical results can be reproduced by others. The framework for data quality assessment will be implemented on top of the DVCS host deployed in WP1 and will access the research data through the DVCS system and write back the results of the different tests into the DVCS system. The framework will also be applicable to external repositories ingested as part of T2.4 in WP2.

[8] `http://librecat.org`

T3.1 *Define metrics for data quality* (M1): Existing metrics for data quality will be discussed and evaluated, based on the case studies. The main criterium for data quality will be its readiness to use by other researchers.

T3.2 *Quality assessment framework* (M6–18): This task will implement the basic framework for data quality assessment and define the architecture that allows various data quality checks to be implemented as services.

T3.3 *Quality assessment services* (M12–24): A number of domain-specific and domain-independent quality assessment services will be implemented that can evaluate data according to the data quality metrics defined in T3.1 to support the pilots defined in WP1.

T3.4 *Integrate validation scripts into platform*: Validation scripts will be created as plug-ins into the platform. In this way it will be possible to easily add additional validation procedures, possibly by third parties or community-created.

T3.5 *Continuous integration*: This work package will investigate as a proof of concept how executable papers can be kept functional across system updates.

WP4: Collaborative Data Access and Reproducibility Workflow System

This work package will combine the version control system and the validation services into a collaborative platform that assists researchers in the process of achieving the status of analytical reproducibility. The system will provide a web interface to give researchers constant and immediate feedback (including notifications using e-mail or social media) on the quality of their data and steps to fix issues that would prevent reproducibility. The results of the validation as well as other social measures will be publicly available by means of badges or other mechanisms, demonstrating quality to all users.

T4.1 *Reproducible paper workflow*: We will define a workflow by which data and the software required for the analysis can be hosted together, and in combination with our use case partners investigate specific customizations of this workflow. Furthermore, we will study the possibility of allowing this workflow to be portable such that it can be automatically executed creating an “executable paper”.

T4.2 *Web interface to data and quality results*: We will extend existing software for hosting DVCS on the Web in order to better support visualizing the data formats used by researchers as well as to display the quality assessment results. In addition, we will automatically extract metadata and publish it using standards such as RDF and Dublin Core.

T4.3 *Social validation of reproducibility*: To enable and even encourage researchers to report reproduction of scientific results, we will add social features to our web interface to show the evaluation of results by users or user actions (e. g., downloads of a dataset).

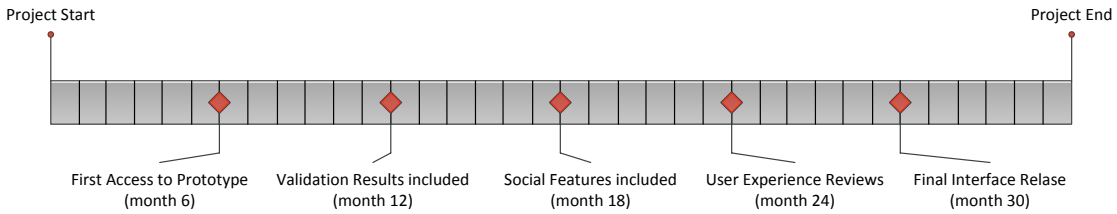

Figure 6: timeline depicting development of the quality assessment and workflow support

WP5: Dissemination and Exploitation

The main objective of this work package is to maximize the applicability and impact of the project results. To this aim, we will liase with relevant networks and initiatives (e. g. COAR and DINI on research data, research information, and institutional repository interoperability; Open Science D-A-CH on licensing; SAFE-PLN on long-term preservation; OpenAIRE on linking data and publications; Research Data Alliance on long tail research data and data citation). Furthermore, we will attract interested parties for integration with institutional repository platforms and research infrastructures. The project will continuously document its results in an open fashion and releases software as free and open-source software.

A timeline depicting the dissemination of the project is shown in Figure 7.

T5.1 Developer challenge coordination **(M12, M24)**: The prototypes and workflow system (WP1, WP4) will be presented to a wider community. To engage with the international community, the project will apply for developer challenges organized as satellite workshops of relevant international conferences (e. g. Semantic Web in Libraries (SWIB), European Library Automation Group (ELAG), Open Repositories, Open Knowledge Festival). Two intended workshops will serve for the refinement of the requirements on the components and to meet interoperability with other repository platforms and research infrastructures.

T5.2 Final Workshop – Implementing Open Science practices in academic institutions **(M36)**: The final project results and its adoption at other academic institutions on the levels of infrastructure, concepts and policies will be discussed with researchers, decision makers and repository administrators.

T5.3 RDA collaboration **(M1–M36)**: The project's framework focuses on issues of data generated in the empirical and experimental sciences. The activities in the areas of best practices for managing long tail research data and data citation will be positioned and synchronized with two international Research Data Alliance (RDA) groups: the "Long Tail of Research Data Interest Group", considers the role of institutional repositories and libraries, and the "Data Citation Working Group", which discusses approaches for dynamic citation of datasets.

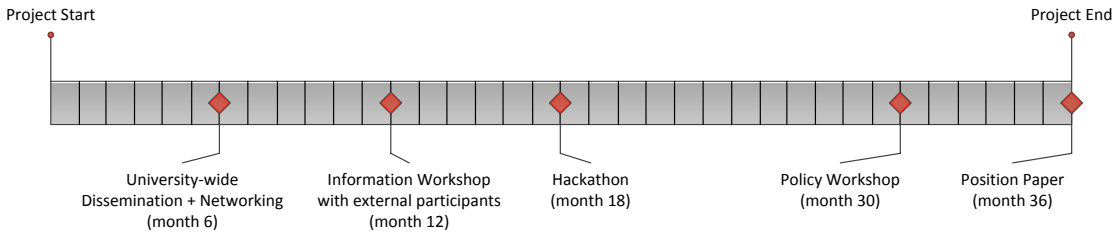

Figure 7: timeline depicting the dissemination of the project

2.4 Measures to meet funding requirements and handle project results

This is an essential project to enhance Bielefeld University's networked information infrastructure. Therefore, significant contributions for sustaining institution-wide policies, tools and services for successful research data handling will be made in the form of personnel and other costs for the following tasks:

- **Policy and advocacy**: Analytical reproducibility as guiding principle of Open Science will provide new opportunities for university management to provide advocacy and incentives for sustainable data sharing. The aim of this proposal is to bring together experiences gained for further policy development.

In close consultation with the focus group, we will harmonize policies for handling research data at the campus with university-level institutions and domain-specific research data infrastructure facilities. Emerging requirements will be incorporated in the university-wide Principles and Guidelines on handling research data at Bielefeld University and will be promoted throughout the university.

During the policy development process, we will prepare and publish recommendations to reflect management of research data as a key strategic challenge for German university management.

- **Integration planning process**: After first requirements of our case studies are met, the integration planning process for enhancing the institutional repository ecosystem will be carried out. In collaboration with the Data Protection Commissioner, IT Security Officer, Chief Information Officer, and Library Directorate, it will target the following legal and organizational measures:
  - In accordance with German data protection acts, a public procedures index will be prepared for co-determination of employees.
  - To improve re-use, the university code of practice for knowledge transfer activities will be considered. In particular, this includes legal instruments such as free and open-source software licences, as well as open licences for databases and creative works.
  - Data storage and backup are key for ensuring the availability of individual projects. Governance models for assigning storage quota need to be developed and implemented. Data storage and backup will be provided by Bielefeld University Computing Center (HRZ).

  The project will help all stakeholders (university management, administration, library, computing center and researchers) to gain a common understanding on requirements toward analytical reproducibility.
- **Long-term preservation for reproducible research**: Bielefeld University is a launch partner of the international SAFE Archiving Federation Private LOCKSS Network (SAFE-PLN) together with Belgian and Canadian universities.[9] SAFE-PLN performs bitstream preservation to ensure the long-term technical stability of born-digital open-access collections at the institutions. Partners have agreed to extend the scope towards content preservation. To this end, it is intended to reuse the proposed continuous integration environment in order to allow for technical re-usability of research data and scientific code.

According to the project proposal obligations, we will make our findings available to the public under open licenses. To this end, we will resort to the Open Access and research data management services already available at Bielefeld University and maintained by Bielefeld University Library. The institutional repository “PUB – Publikationen an der Universität Bielefeld” allows for self-archiving of publications and research data according to national (DINI) and international standards (OpenAIRE, DataCite, OAI-PMH). To increase the visibility of project results, software distributions will be additionally released through open source software hosts such GitHub and ScourceForge.

Publication of project findings will take the interests of persons into account and be based on open licenses.

[9] `http://www.safepln.org/`

3 Bibliography

[1] Paul Ayris, Pablo Achard, Serge Fdida, Stefan Gradmann, Wolfram Horstmann, Ignasi Labastida, Liz Lyon, Katrien Maes, Susan Reilly, and Anja Smit. LERU roadmap for research data, 2013.

[2] Grady Booch. *Object Oriented Design: With Applications*. The Benjamin/Cummings Series in Ada and Software Engineering. Benjamin/Cummings, 1991.

[3] CITEC. CITEC Open Science Manifesto. `http://www.cit-ec.de/openscience/manifesto`, 2013.

[4] Deutsche Forschungsgemeinschaft (DFG). Recommendations for secure storage and availability of digital primary research data. `http://dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901_en.pdf`, 2009.

[5] Stefan Friedhoff, Christian Meier zu Verl, Christian Pietsch, Christian Meyer, Johanna Vompras, and Stefan Liebig. Replicability and comprehensibility of social research and its technical implementation. *RatSWD Working Paper Series*, 219, 2013.

[6] Thomas Herndon, Michael Ash, and Robert Pollin. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, 2013.

[7] Konrad Hinsen. A data and code model for reproducible research and executable papers. In *Proceedings of the International Conference on Computational Science (ICCS)*, pages 579–588, 2011.

[8] L. N. Joppa, G. McInerny, R. Harper, L. Salido, K. Takeda, K. O'Hara, D. Gavaghan, and S. Emmott. Troubling trends in scientific software use. *Science*, 340(6134):814–815, may 2013.

[9] Tomi Kauppinen and Giovana Mira de Espindola. Linked Open Science – communicating, sharing and evaluating data, methods and results for executable papers. In *Proceedings of the International Conference on Computational Science (ICCS)*, pages 1–6, 2011.

[10] Stefan Kramer, Amber Leahey, Humphrey Southall, Johanna Vompras, and Joachim Wackerow. Using RDF to describe and link social science data to related resources on the Web, 2012.

[11] Florian Lier, Frederic Siepmann, Thilo Paul-Stueve, Sebastian Wrede, Sven Wachsmuth, and Ingo Lütkebohle. Facilitating research cooperation through linking and sharing of heterogenous research artifacts. In Harald Sack and Tassilo Pellegrini, editors, *Proceedings of the 8th International Conference on Semantic Systems (I-SEMANTICS '12)*, pages 157–164. ACM, 2012.

[12] Florian Lier, Johannes Wienke, Arne Nordmann, Sven Wachsmuth, and Sebastian Wrede. The Cognitive Interaction Toolkit – improving reproducibility of robotic systems experiments. In Davide Brugali, editor, *SIMPAR 2014, LNAI*, pages 400–411. Springer International Publishing Switzerland, 2014.

[13] John McCrae, Cord Wiljes, and Philipp Cimiano. Towards assured data quality and validation by data certification. *1st Workshop on Linked Data Quality (LDQ2014) (co-located with SEMANTiCS 2014)*, 2014.

[14] Organisation for Economic Co-operation and Development (OECD). OECD Principles and Guidelines for Access to Research Data from Public Funding. Technical report, Organisation for Economic Co-operation and Development, 2007.

[15] Roger D. Peng. Reproducible research and biostatistics. *Biostatistics (Oxford, England)*, 10(3):405–8, July 2009.

[16] Karl R. Popper. *The Logic of Scientific Discovery*. Hutchinson & Co, New York, 1959. First published in German as *Logik der Forschung*, Wien, 1934.

[17] Karthik Ram. Git can facilitate greater reproducibility and increased transparency in science. *Source Code for Biology and Medicine*, 8(1):7, 2013.

[18] Carmen M. Reinhart and Kenneth S. Rogoff. Growth in a time of debt. Working Paper 15639, National Bureau of Economic Research, January 2010.

[19] Victoria Stodden, Peixuan Guo, and Zhaokun Ma. Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLoS ONE*, 8(6):e67111, jun 2013.

[20] Johanna Vompras, Jochen Schirrwagen, and Wolfram Horstmann. Die Bibliothek als Dienstleister für den Umgang mit Forschungsdaten. In Silke Schomburg, Claus Leggewie, Henning Lobin, and Cornelius Puschmann, editors, *Digitale Wissenschaft: Stand und Entwicklung digital vernetzter Forschung in Deutschland*, pages 101–106. hbz, 2011.

[21] Cord Wiljes and Philipp Cimiano. Linked data for the natural sciences: Two use cases in chemistry and biology. In *Proceedings of the Workshop on the Semantic Publishing (SePublica 2012)*, pages 48–59, 2012.

[22] Cord Wiljes, Najko Jahn, Florian Lier, Thilo Paul-Stueve, Johanna Vompras, Christian Pietsch, and Philipp Cimiano. Towards linked research data: An institutional approach. In Alexander García Castro, Christoph Lange, Phillip Lord, and Robert Stevens, editors, *3rd Workshop on Semantic Publishing (SePublica)*, pages 27–38, 2013.