8th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2015 [cit. 2015-12-15]. Available from: http://nrgl.techlib.cz/index.php/Proceedings. ISSN 2336-5021.

REPOSITORY WORKFLOW FOR INTERLINKING DATA WITH GREY LITERATURE

Johanna Vompras, Jochen Schirrwagen

johanna.vompras@uni-bielelefeld.de, jochen.schirrwagen@uni-bielefeld.de

Bielefeld University Library, Universitätsstr. 25, Bielefeld, 33615, Germany

Abstract

Publishing data is more and more considered as part of the research process. While funder mandates and journal policies demand the disclosure of research data at the time of article publication there is still a lack of guidelines and workflows to reference data from grey literature. Based on multidisciplinary examples found in our repository ’PUB’ we present a user friendly generalized framework for interlinking research data with grey literature. This way, we are not only increasing the number of ’grey’ non-textual research outputs - including research data - but also foster awareness of its sharing and re-use in scientific communities.

Keywords

Research Data, Linking Data to Publications, RDM, Institutional Research Infrastructures

Introduction

Research data and other nontraditional output types are increasingly considered as valuable as research papers and other textual publications. Research data that can be persistently identified relate to increased visibility, citability and re-use [4]. They may allow for reproducibility and validation of research results. This imposes requirements on the research data management, documentation using descriptive and disciplinary metadata standards, storage, interlinking with their publication, curation and optional anonymization.

Grey literature mostly describes how data was collected, generated, or processed. It is not only characterized by a large heterogeneity of publication types (dissertations, technical reports, data handbooks, working papers, methods reports, newsletters and bulletins etc.) but also by the question how research data is referenced therein. We analyzed grey literature in our repository PUB, in particular publications with supplementary material or linked to data sets. We found various types of supplementary material, among them research data, software and other kinds of research output that can be seen as discrete resources.

Aims

We propose certain measurements to improve publishing of research data and other research output linked to the underlying grey literature resulting in a better visibility and discovery of all research results. We see the following key requirements:

- **Citation:** authors should cite datasets in their underlying publications (enabling data location and validation by the reader) and vice versa (enabling the reader to understand context and methodology).
- **Awareness:** author guidelines for writing grey literature material should recommend the registration and deposit of research data if generated or used in the research process; design and implementation of corresponding publication workflows [2] need to be oriented towards intuitive and efficient user interfaces to minimize additional effort.
- **Reproducibility of Research:** publications should be complemented by any documents supporting interpretation and replication of the research data and helping to give insight into the resulting research findings.
- **Publication of Research Data:** research data (either created or re-used) that is needed to validate research results should be prepared for deposit and archiving.

State of the Art

Several funding bodies, as well as policy makers and research councils increasingly propagate that publicly funded research data should be openly available to the scientific community. The most German funders (like DFG[1]) and funders from the international landscape follow this guidance and encourage researchers to share and contextualize their data and research output. Through data disclosure policies formulated by journals (e.g. Nature, PLOS One) data, materials, codes, or scripts forming the basis for the respective publications have to be deposited within a repository at publication time.

In a meanwhile, policies imposed by funding agencies received wide recognition, but there is still a lack of institutional data policies. For example, guidelines on handling research data might be defined either at multiple administrative levels or they might not be anchored in the institutional policy at all. In addition, their liability might be misunderstood by researchers,

[1] DFG: German Research Foundation: **http://www.dfg.de/en/**

or there might be a lack of technical and organizational support to fulfill them, e.g. by missing research data management services or library services supporting registration and dissemination, and contextualization of research data. In Germany, the awareness of the professional handling on research data has already reached the universities. In May 2014, the General Meeting of the German Rectors' Conference (HRK) recommended the management of research data to be a strategic function of university management and calls the Universities upon to create the structural framework for efficient research data management for the whole institution.

Research Data as an integral part of grey literature

To illustrate the variety of publications with related or linked research results we picked some samples to showcase the current situation.

- *Enhanced Publication:* In 2011, a PhD student at TU Delft published his dissertation in the institutional repository and linked to the associated datasets published in the 3TU.Datacentrum[2].
- *Technical Report of the Collaborative Research Center 882 (SFB882)*[3]: The technical report itself contains parts which might be considered as research data (e.g. questionnaire, codebook in Appendix). The data is embedded within the document in a non-standardized way. Further metadata details on the registered data can be found as a DOI reference (here: http://dx.doi.org/10.4119/unibi/sfb882.2014.12), which links from the PDF to the data landing page.
- *Bielefeld Working Paper in "Economics and Management"*[4]: The landing page and metadata of the working paper relate to the source code and links to a stand-alone "research data set" http://doi.org/10.4119/unibi/2674041 which is in turn cited by other publications.
- *Software Publication:* Connecting version control systems, e.g. GitHub, with repositories, e.g. zenodo, allows for archiving of software snapshots or releases with persistent identifiers and thus makes code citable and put it into context with related publications and projects [3].

While publishing research data with the underlying publication by the authors themselves is a fundamental first step, services on top are needed which allow for aggregation, linking, knowledge extraction and discovery of those research artifacts across heterogeneous data sources. There are a number of initiatives that are striving to achieve such aims:

- *Data Literature Interlinking (DLI) service*[5]: The DLI service is a result of a collaboration between data centres, publishers, and research organizations that provides 'authoritative' links between datasets and underlying research literature.
- *InFOLiS I:* The aim of this DFG-funded project [8] was the development of techniques to discover links between publications and research data (in the Social Sciences) automatically and to retrospectively integrate them into the current retrieval systems.

[2] **http://t1p.de/ep-dissertation-tudelft**

[3] **http://pub.uni-bielefeld.de/publication/2730392**

[4] **http://pub.uni-bielefeld.de/publication/2723277**

[5] **http://dliservice.research-infrastructures.eu/**

- *OpenAIRE:* Aggregation and knowledge extraction from literature and data repositories are key components of the OpenAIRE scholarly communication infrastructure[6] that allows for contextualization of research results [9].
- *Research Data Switchboard:* By aggregation of links between datasets, publications and projects across different registries this service allows for discovery of datasets and related information[7].
- *DataCite Metadata Search*: This service[8] allows finding research data but also grey literature that has been minted with a DataCite DOI. In DataCite metadata relations to other resources can be stated, e.g. a dataset that relates with its underlying publication.
- *E-Science Funding Programme on the federal state level (Baden-Württemberg)*: The project bwDataDiss[9] develops an interdisciplinary infrastructure for describing, storing and linking research data with electronic thesis.

Standardized Relations in Metadata

The relation between the data and underlying publication needs to be manifested in metadata. One example is the DataCite metadata schema[10] that allows describing related resources of e.g. a research dataset using the element 'relatedIdentifier'.

```
<relatedIdentifier
  relatedIdentifierType="DOI"
  relationType="IsCitedBy">
  10.1234/fooBar
</relatedIdentifier>
```

In contrast, there is a variety of metadata standards to describe electronic theses and dissertations[11], making interoperability a challenging task. XMetaDissPlus is the metadata standard used in Germany to describe and transfer electronic dissertations [6]. Using Dublin Core elements the standard provides basic support for tagging relations to other resources that are part of a dissertation like research data.

```
<dc:relation xsi:type="dcterms:URI">http://dx.doi.org/10.1234/fooBar
</dc:relation>
<dcterms:requires xsi:type="dcterms:URI">http://dx.doi.org/10.1234/fooBar
</dcterms:requires>
```

Addressing Challenges

Grey literature often lacks standards for its creation, publication and distribution. Its target application and target audience may also vary, e.g. it may be shared only among a researcher group and never be published to the public. Grey literature is therefore often difficult to discover, access, and evaluate. Furthermore, it lacks of guidelines and good practices about research data linkage in grey literature. However, it offers opportunities for research libraries

[6] **https://www.openaire.eu**

[7] **http://www.rd-switchboard.org/**

[8] **https://www.datacite.org/services/find-dataset.html**

[9] **http://www.alwr-bw.de/kooperationen/bwdatadiss (in German)**

[10] **http://schema.datacite.org/**

[11] **http://www.ndltd.org/standards/metadata**

to contribute with services for research data management at institutional level [7]. In particular we encounter the following challenges:

- Data related to a publication (e.g. primary data, software products, or scripts used for data analysis) is mostly just mentioned in footnotes or bibliographies without a proper citation. Thus, this data is not searchable or accessible through public research data catalogues and academic search engines.
- Such discrete resources need to be registered and mutually linked with the grey literature document as shown in Figure 1.
- Authoring tools are needed that support guidelines for e.g. the production of scientific and technical reports [11] and support standards for citing data.
- Institutional policies – if they exist – often do not cover grey literature and research data as a regular component of the research output.
- Solutions for the Research Data Management (RDM) should be generic enough, that they can be applied to many disciplines.
- Libraries can take the role to promote awareness to researchers about a possible publication of research data and its sharing.

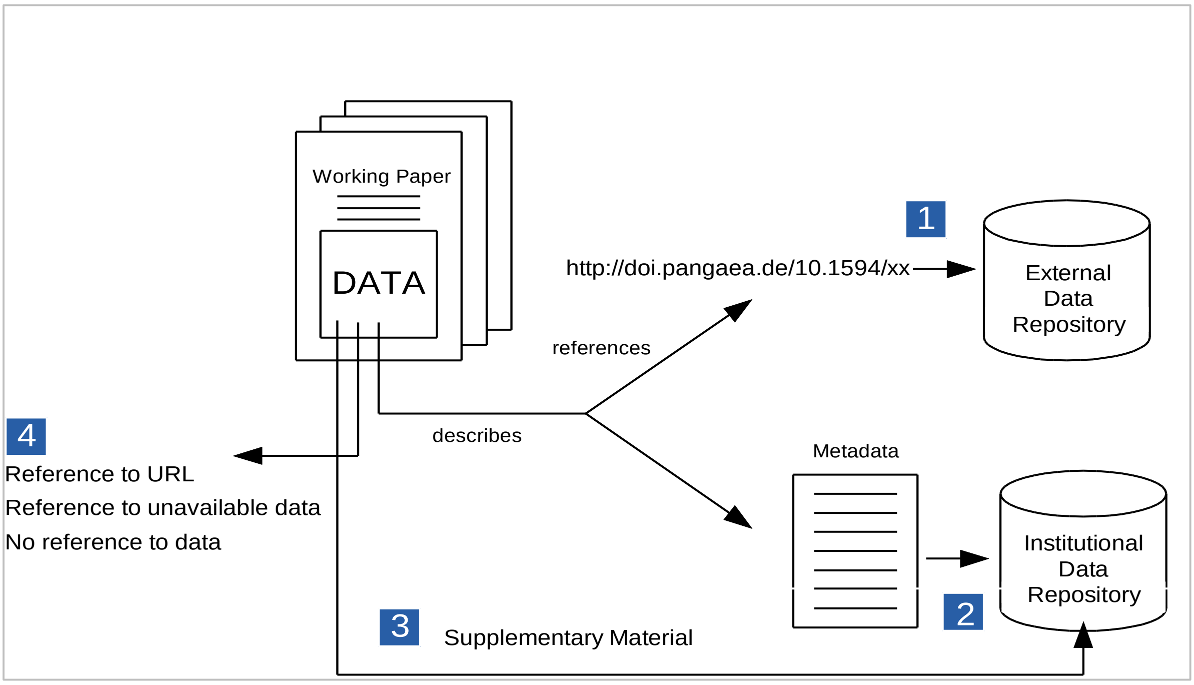


Figure 1 Use Cases for Linking Grey Literature with Data

Institutional Context with PUB

In 2009, the *Rektorat* of Bielefeld University launched the INFORMIUM[12] Initiative in order to identify the requirements for a university wide and generic research data infrastructure. As a result, Bielefeld University has passed guidelines and policies on research data management, being the first German university to form an institution-wide agreement upon

[12] https://data.uni-bielefeld.de/en/informium

standards for research data handling among all stakeholders. The university-wide policy calls on researchers

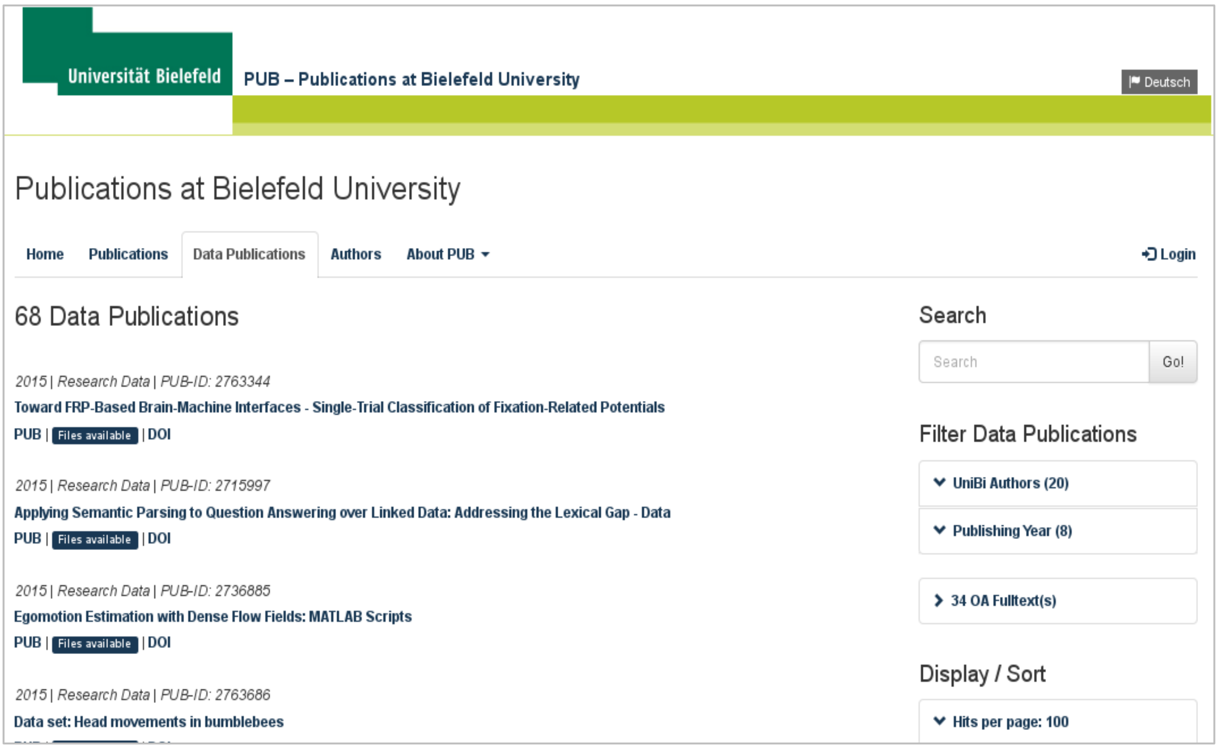
> (i) to take advantage of the university's advisory services for research data management and
> (ii) to publish research data through registered research data repositories[13]. To this end, Bielefeld University Library offers comprehensive advisory services on data management planning, data publication and preservation through its institutional repository PUB – Publications at Bielefeld University[14].

"PUB – Publications at Bielefeld University" is used to reflect the work of the university's researchers. Through easy-to-handle embedding possibilities, different views on the data (e.g publications of a single researcher, research group, department, or whole faculty) can be visualized as lists – optionally with faceted search – on the respective institution's websites. PUB is a hybrid institutional repository depositing and disseminating data and publications. The repository is compatible with OpenAIRE which supports and monitors the Open Access mandate in EU Horizon 2020.

Technically, PUB is based on the LibreCat[15] framework developed by the university libraries Gent, Lund and Bielefeld. Through its data processing routines for data oriented applications it facilitates the normalization of metadata and provides plugins for import and export. PUB is well integrated in the international scholarly communication infrastructure, e.g. by importing from large bibliographic databases and thematic repositories. It supports established machine interfaces (OAI-PMH, SRU, CQL) and metadata formats (Dublin Core, DataCite Metadata Kernel, MODS, XMetaDissPlus) to serve aggregative services like BASE[16], OpenAIRE, DataCite, EuropePMC[17] and DNB[18].

To ensure failure safety and reliability, PUB participates in the international distributed preservation repository network, SAFE Private LOCKSS Network[19], with the aim to preserve digital objects for future generations and to minimize the risk of data loss – caused by hardware breakdowns, obsolescence or natural disasters, or even human errors – over the long-term. The overall idea of SAFE-PLN is to make multiple copies (here: seven) as preservation strategy and to disseminate these copies throughout the world, in places considered to be safe. In the event of an unfortunate loss, data can be restored from one of the other preservation nodes, which all act in an autonomous and independent way at both financial and administrative level.

[13] **http://re3data.org**
[14] **http://www.pub.uni-bielefeld.de**
[15] **http://librecat.org**
[16] **http://base-search.net**
[17] **http://europepmc.org/LabsLink**
[18] **http://www.dnb.de/DE/Wir/Kooperation/dissonline/dissonline_node.html**
[19] **http://www.safepln.org**



Publications at Bielefeld University

68 Data Publications

2015 | Research Data | PUB-ID: 2763344
Toward FRP-Based Brain-Machine Interfaces - Single-Trial Classification of Fixation-Related Potentials
PUB | Files available | DOI

2015 | Research Data | PUB-ID: 2715997
Applying Semantic Parsing to Question Answering over Linked Data: Addressing the Lexical Gap - Data
PUB | Files available | DOI

2015 | Research Data | PUB-ID: 2736885
Egomotion Estimation with Dense Flow Fields: MATLAB Scripts
PUB | Files available | DOI

2015 | Research Data | PUB-ID: 2763686
Data set: Head movements in bumblebees

Figure 2 PUB Search Interface for Research Data

PUB has been extended by several aspects of data contextualization in the course of the introduction of the institutional research data policy[20] in 2013. As one measure Bielefeld University qualified as a publication agency for DataCite DOI. In PUB, the DOI registration of research data is part of the publication process. The persistent identification makes sure that the data stays available unchanged over time for later verification and re-use. Thus, the DOI can be used to cite the data in the manuscript. In general, the DOI resolves to a landing page in PUB, except for bilateral agreements made with research groups where the DOI may resolve to databases at their research institute. Research data in PUB have their own view as shown in Figure 2. They are part of more than 46,300 bibliographic references and publications in PUB. Thus, PUB provides increased access to a rich array of research output – among them grey literature and research data – as summarized in Table 1.

Quantity	**Publication Type**	**With Suppl.Material**
1670	thesis / dissertation	18
814	conference proceeding / paper	47
656	working paper	1
120	report	4
74	research data	
18	preprint	0

Table 1 Distribution of Open Access Grey Literature and Research Data in PUB

[20] https:ttps://data.uni-bielefeld.de/en/resolution

Some of these publications have supplementary material attached or data embedded in the publication, including research data (tabular data files, questionnaires and variables), software, algorithms, and multimedia files. This motivates for the design and implementation of a workflow for the discrete publication of literature, data and software.

Conceptual Design

In order to develop a common, discipline-independent model for interlinking grey literature with data, we have first analyzed concrete examples available in PUB to find out how research data is put into context. We found discipline-specific aspects based on differences in the particular research process and diversity in interpretation of the data life cycle. In addition, the maturity and the stages of how data citation principles are applied among disciplines are deeply divided. For example, in the Social Sciences, it is quite common to cite external data (e.g. census data), which have been used to respond the own research question. Conceptually, we consider the following cases in PUB:

A Classical supplementary material attached to the main manuscript: Any material (tables, figures, descriptions) that provides additional information and is published together with the main manuscript so that it depends on it and is not a discrete publication by itself.

B Stand-alone "research data": A dataset, software etc. that was generated, implemented, analyzed or otherwise used in the context of a research question and therefore constitutes a discrete resource. It benefits from referencing to a publication and can be cited in other works and gives credit to its authors or contributors.

C Links to external data: References to data either used to derive findings for the own research (e.g. any third party data) or data generated during the research (e.g. raw data, processed data, results, etc.). The data has been published or archived externally in a discipline specific repository.

In case researchers consider publishing their “own” data, the Library Services might support them in the decision making process. For example, questions about any restrictions placed on sharing the data (e.g. ethical, commercial, protection of personal data, intellectual property) are discussed.

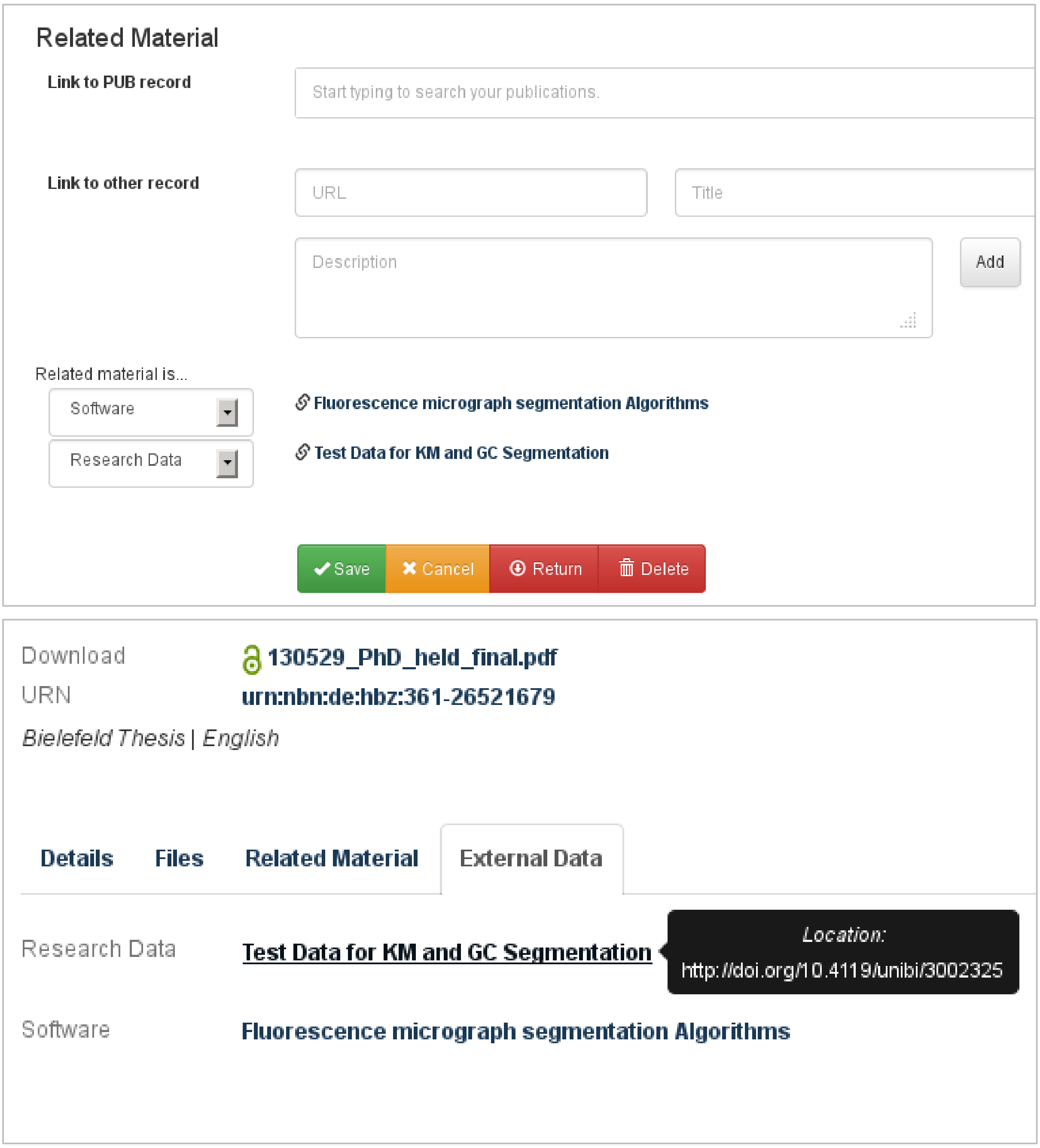
Implementation

Advisory Support for Publication

Another aspect of supporting researchers in publishing their data within grey literature (e.g. PhD thesis) is to incorporate data publication workflows to the thesis submission process. After a successful thesis defense, researchers have to accomplish a publication of their thesis – as part of the graduation requirements – either in printed form (e.g. book) or online through the University’s Library publication services (pdf-file). Since the latter is the most frequently used choice, we are building on this point to sensitize the PhD candidates to publish their data attached to the printed thesis (e.g. on storage media) or enquire about the existence of data worth being published or reused within a broader community.

Thesis Submission Process

The PUB system supports a general user interface to define relations between resources of different type. A dissertation can be associated with both an already existing and registered research data in PUB and any kind of external resources, like deposited in a disciplinary repository. It is also possible to make plain references to software sources in version control systems like GitHub[21].



Related Material

Link to PUB record: Start typing to search your publications.

Link to other record: URL | Title | Description

Related material is...

Software — Fluorescence micrograph segmentation Algorithms

Research Data — Test Data for KM and GC Segmentation

Download: 130529_PhD_held_final.pdf

URN: urn:nbn:de:hbz:361-26521679

Bielefeld Thesis | English

Research Data: Test Data for KM and GC Segmentation

Location: http://doi.org/10.4119/unibi/3002325

Software: Fluorescence micrograph segmentation Algorithms

Figure 3 Submitting a PhD-Thesis with References to Research Data (a) and the corresponding view in the Landing Page (b)

Figure 3 (b) shows the respective landing page of the publication with the visualized interlinked data which is public to all users. Here, all of the interlinked resources are represented

[21] https://github.com

by hyperlinks where the target depends on the relation and resource type (see column "example" in Table 2).

	Relations	**Identifier**	**Resource type**	**Target example**
Related material	suppl. Materials	PUB-ID, FILE-ID	any	any uploaded file
Publications in PUB	is-part-of, earlier-version, cites, is-cited-by ...	PUB-ID	research data (as stand-alone publication)	PUB landing page
External Data	is-part-of, earlier-version, cites, is-cited-by, uses ...	DOI, URL	software, data sets	link to GitHub, Dryad, biological databases

Table 2 Possibilities of relating data to a PhD thesis in PUB

6. Conclusion and Future Work

In this paper we discussed challenges of linking research data with grey literature and presented an organizational and technical workflow that enables its publication, inter-linking and preservation through our institutional repository. By implementing and promoting institutional research data policy, technical infrastructure and organizational support to our researchers we increase their awareness of possibilities to publish and share data. In the context of the just started DFG funded CONQUAIRE project [5] interdisciplinary research teams, the CITEC Semantic Computing Group[22] and the University Library collaborate towards an infrastructure supporting analytical reproducibility of scientific data tightly integrated in the research process.

References

[1] BRINEY, Kristin, Abigail GOBEN, and Lisa ZILINSKI. Do You Have an Institutional Data Policy? A Review of the Current Landscape of Library Data Services and Institutional Data Policies. *Journal of Librarianship and Scholarly Communication* [online]. Pacific University Libraries, 2015, **3**(2). E-ISSN: 2162-3309 Available from: **http://jlsc-pub.org/articles/abstract/10.7710/2162-3309.1232/**.

[2] BLOOM, Theodora et al. *Workflows for Research Data Publishing: Models and Key Components (Submitted Version)*. 2015. DOI: **10.5281/zenodo.20308**.

[3] POTTER, M. and T. SMITH. Making code citable with Zenodo and GitHub. In: *Software Sustainibility Institute* [online]. 2015. Available from: **http://www.software.ac.uk/node/1720**.

[4] PIWOWAR H. A and T.J. VISION. Data reuse and the open data citation advantage. *PeerJ* [online]. PeerJ, 2013. ISSN: 2167-8359. Available from: **https://dx.doi.org/10.7717/peerj.175**.

[22] **http://www.sc.cit-ec.uni-bielefeld.de/**

[5] CIMIANO, P. et al. *CONQUAIRE: Continuous quality control for research data to ensure reproducibility: an institutional approach*. 2015. DOI: 10.5281/zenodo.31298. Project Proposal to Deutsche Forschungsgemeinschaft in the Programme.

[6] KOORDINIERUNGSSTELLE DISSONLINE (Ed.). *Referenzbeschreibung XMetaDissPlus v2.2*. Leipzig: Deutsche Nationalbibliothek, 2012. Available from: **http://nbn-resolving.de/urn:nbn:de:101-2012022107**.

[7] AYRIS, P., P. ACHARD and S. FDIDA, et al. *LERU Roadmap for Research Data.* LERU Advice Paper. Leuven: LERU, 2013, vol 14. LERU. Available from: **http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf**.

[8] BOLAND, K., D. RITZE, K. ECKERT and B. MATHIAK. Identifying references to datasets in publications. In: *Theory and Practice of Digital Libraries*. Berlin: Springer, 2012, p. 150-161. ISBN: 978-3-642-33289-0. ISBN 978-3-642-33290-6. DOI: 10.1007/978-3-642-33290-6.

[9] KOBOS, M., Ł. BOLIKOWSKI, M. HORST, P. MANGHI, N. MANOLA, & J. SCHWIRRWATEN. Information Inference in Scholarly Communication Infrastructures: The OpenAIREplus Project Experience. In: *Procedia Computer Science*. Elsevier, 2014, vol. 38, p. 92-99. DOI: 10.1016/j.procs.2014.10.016. ISSN: 1877-0509.

[10] HRK: GERMAN RECTORS' CONFERENCE. *Recommendation of the 16th General Meeting of the HRK, 13 May 2014: Management of research data – a key strategic challenge for university management*. Bonn: HRK, 2014. Available from: **http://www.hrk.de/uploads/tx_szconvention/HRK_Empfehlung_Forschungsdaten_13052014_EN.pdf**.

[11] DE CASTRO, P., & S. SALINETTI. Grey literature: challenges and responsibilities for authors and editors. In: *Science Editors' Handbook*. European Association of Science Editors, 2013. ISBN 978-0-905988-11-5. Available from: **http://www.ease.org.uk/sites/default/files/6-4.pdf**.