Research in Context

Schirrwagen, Jochen ; Jahn, Najko
2013

Seminar on Providing Access to Grey Literature 2013: The 6th year of the seminar focused on storage and providing access to the grey literature, 23th October 2013 [online]. Praha: National Library of Technology, 2013. Available at WWW: http://nusl.techlib.cz/Sborniky. ISSN 1803-6015.

RESEARCH IN CONTEXT (IN VIEW OF RECENT RESULTS FROM OPENAIRE PLUS AND FROM THE LIBRARY PERSPECTIVE)

JOCHEN SCHIRRWAGEN, NAJKO JAHN

jochen.schirrwagen@uni-bielefeld.de, najko.jahn@uni-bielefeld.de

Bielefeld University Library, Bielefeld, Germany

Abstract
Context of scholarly results improves discovery, sharing and re-use in library and research communities. OpenAIREplus facilitates access to European research. Its mission is to interlink research publications, data, contributors and grants. Introducing pilots for the Social & Life Sciences we show how disciplinary services can be used to enhance publications in cross-disciplinary environments. The approach followed originates from text-mining and interlinks innovative repository services and research data infrastructures.

Keyword
Open Access, Open Science, Enhanced Publication, Research Contextualization.

1. MOTIVATION

With the advent of a culture of Open Access and Open Science in conjunction with data-driven science new opportunities arise to publish research results with all available related information. It allows authors to better and more transparently present their findings ensuing from the research question, linked with the arising data, as well as the methodology used and conclusions described in their papers. For example, contributors of research data can be better credited. Scientists interested in verifying or re-using results of their colleagues gain a better insight and can ideally access the actual tools that are linked to the research results. Funding organizations could also be provided with more detailed assessment of the impact of their funded research.

In recent years concepts have been suggested and experimented with to realize such a scenario. From the DRIVER-II project findings, “Enhanced Publications” have been described with respect to modelling of their structural elements, and their support in repository platforms [Woutersen-Windhouwer, 2009]. Similarly the linked open data approach suggests the concept of “Linked Science” by semantic modelling of scientific resources [Kauppinen, 2013]. Elsevier’s “Article of the Future” project has shown how scholarly articles in electronic journals can be linked with underlying research data in data repositories [IJsbrand, 2013].

In this paper we describe and present recent results and developments on how to link and enrich scholarly publications and research data in two complementing environments. In section two, we present our institutional approach at Bielefeld University by using services for publication and research data management. In section three we describe the challenges and

opportunities to enhance and contextualize research output in the cross-disciplinary OpenAIRE infrastructure.

2. Contextualization at the Source – an Institutional Approach

At Bielefeld University, significant parts of the research infrastructure are networked on a university-wide basis. Within this network, the central service for the collection and dissemination of publications is the institutional repository and academic bibliography PUB[1] - Publications at Bielefeld University. Highly integrated in the university-wide IT-infrastructure, PUB reuses Bielefeld University's authentication and authorization infrastructure, staff and department profiles as well as grant information to enrich registered publications. Based on this integration, researchers and departments can create dedicated publication profiles to be embedded in the personal or working group homepages.

One particular concern is to link PUB with external bibliographic resources. This shall allow easy imports and ex post enrichment of external information such as availability of Open Access copies or research database links.

Among those sources referenced are Web of Science, Europe PubMed Central, arXiv, INSPIRE and the Directory of Open Access Journals.

A future requirement is to provide a publishing and archiving service for long-tail research data, i.e. research data that is not in the scope of an already existing disciplinary data infrastructure. This service is embedded at a university-wide strategy towards interlinking research data with publications and research information management. It is accompanied by disciplinary pilots [Friedhoff, 2013].

2.1 Examples

The following examples serve for illustration on how research results in PUB can be enriched with related information and represented in a contextualized form.

Linking with the CITEC Open Research Platform and the Interactive Toolkit

The Center of Excellence Cognitive Interaction Technology (CITEC) is an interdisciplinary research institute at Bielefeld University. CITEC follows an Open Science policy and provides a service infrastructure addressing for instance Open Source software and Open Data databases. The Cognitive Interaction Toolkit [Lier, 2012] enables continuous integration of those research artefacts. Bibliographic metadata is registered in the university-wide service and imported into the Toolkit. Within the toolkit, researches link the metadata with other research artefacts aggregated by the Toolkit. In the end, these enhanced publications are exposed as Linked Data [Wiljes, C., 2013].

[1] PUB – Publications at Bielefeld University: http://pub.uni-bielefeld.de

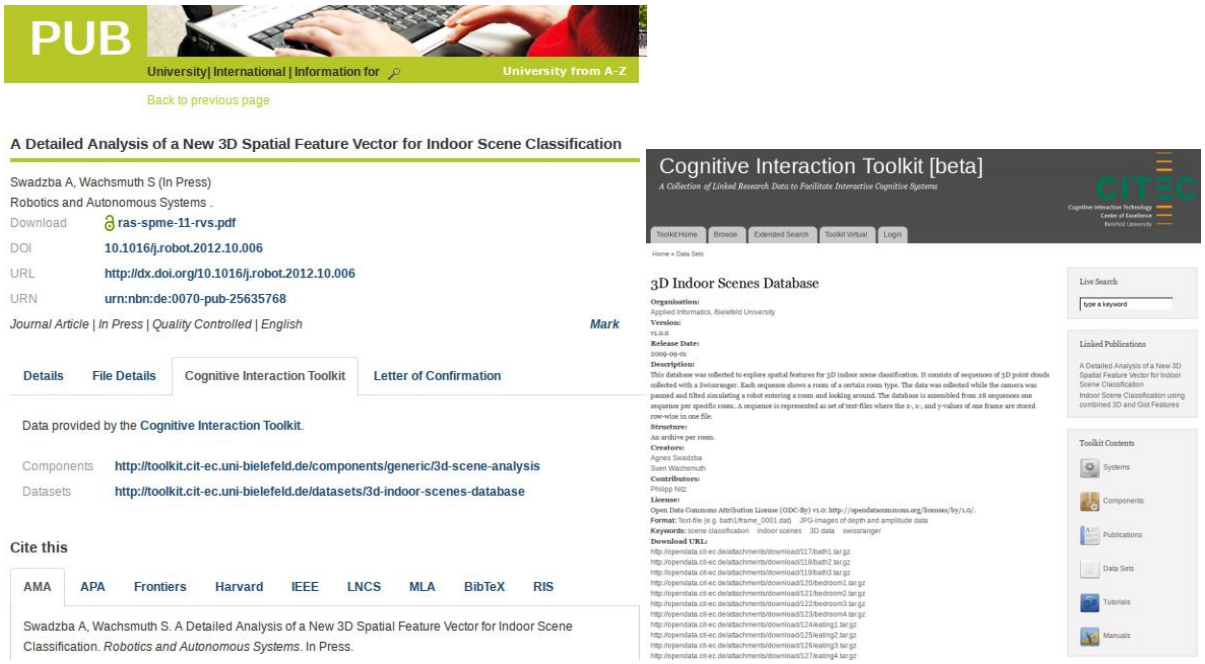

Figure 1 Cross-References between publications (PUB, left) and research datasets (CITEC Interactive Toolkit

Linking with Europe PubMed Central and EBI databases

Europe PubMed Central (EuropePMC) is one of the leading thematic portals in the Life Sciences. Publications in this domain are typically assigned with PMIDs and PMC IDs. As a form of enrichment PUB can link publications with those identifiers to EuropePMC as shown in figure 2.

Figure 2 Publication linkage in PUB with EuropePMC

Recently EuropePMC has developed a value added service the "External Links Service"[2] [Kinsey, 2013]. It is a vehicle to provide further information to publications in Europe PMC. These can be fulltext links, links to databases and datasets or even educational coursework. During the summer of 2013 PUB provided links to more than 250 Open Access publications[3] otherwise not yet available in Europe PMC. Conversely PUB links from its publications to research data at EBI automatically. Up to now more than 100.000 links to bio entities have been determined and are visible at Bielefeld University website on article level.

3. Linking Data, Publications and other Scholarly Information in OpenAIRE

The linking and contextualisation of research results at the infrastructure level is far more challenging than it is at the local level.

OpenAIRE[4] – the Open Access Infrastructure for Research in Europe – is building its 2nd generation participatory infrastructure. Proceeded by the European Commission Open Access pilot which collects FP7 and ERC funded research publications, it now provides access to the research results of European funded projects, extends to other national funders and includes Open Access publications aggregated from the DRIVER network of institutional and thematic repositories. Moreover, by collecting information on research datasets the revised infrastructure includes methods to interlink related entities of scholarly communication. This provides the user with a rich information graph which can navigate between publications, and if available research data, research information and statistics.

However OpenAIRE is faced with major challenges. These issues are usually because of the sheer diversity of scientific data, and the subject-specific infrastructure interfaces. Cultural practices for managing data and publications in subject communities also differ greatly. Therefore OpenAIRE closely works with scientific partners from different scientific fields: DANS[5] for the Social Sciences & Humanities, EMBL-EBI[6] for the Life Sciences and BADC[7] for earth observation data.

OpenAIRE's network of 33-pan European advocacy offices liaise with initiatives (COAR[8], LIBER[9], RDA[10]) to promote the benefits of Open Access to data and publications, to provide

---

[2] External Links Service: http://europepmc.org/LabsLink

[3] http://europepmc.org/search?page=1&query=LABS_PUBS:1056&sortby=Relevance

[4] OpenAIRE is available at http://www.openaire.eu and with its new infrastructure located at http://beta.openaire.eu

[5] Data Archiving and Networked Services: http://www.dans.knaw.nl

[6] European Bioinformatics Institute: http://www.ebi.ac.uk

[7] British Atmospheric Data Centre: http://badc.nerc.ac.uk

[8] Confederation of Open Access Repositories: http://www.coar-repositories.org/

[9] Association of European Research Libraries: http://www.libereurope.eu/

guidance for researchers and managers of participating data providers. This is furthered by the OpenAIRE guidelines[11] which assist repository managers of the following typologies to encode links from publications to projects and research data in the metadata:

- Scholarly publications in institutional and thematic repositories using OAI-PMH and usage guidelines of OAI-DC metadata
- Research data in data repositories adapting the DataCite metadata kernel [Starr, 2011]
- Research information about funded projects in CRIS applications

3.1. Subject specific Pilots for Enhanced Publications

OpenAIRE aims to build a cross-disciplinary infrastructure that represents publications linked with underlying research data. From the start, subject-specific practices of researchers and the management of publication and research data linkage in their communities needs to be fully understood. For this reason two demonstrators have been constructed to showcase the interlinking of data and publications within the Life Sciences and Social Sciences & Humanities. They address different aspects and serve for the identification of issues and current practices, and to stimulate the development of a cross-disciplinary model in the OpenAIRE infrastructure [Hoogerwerf, 2013].

Life Sciences demonstrator

The Life Sciences demonstrator[12] is based on publications originating in Europe PMC which actively links to biological research data. These publications shall be “re-used” in a generic aggregative infrastructure that can provide further information related to these publications (e.g. project information, similar publications, citation information). However it was identified that the lack of a standardized exchange format that encode such enriched bibliographic metadata and the use of domain specific API’s for data access impedes its transfer among infrastructures.

The demonstrator shows publications that were identified by PMIDs in the Bielefeld PUB service as described above. The demonstrator relates the publications with other entities, namely authors, bio-entities and FP7 project information in a graphical user interface of a web application, see figure 3.

The enrichments are captured automatically. For a given PMID the EuropePMC RESTful service[13] can be queried for linked information on:

[10] Long Tail of Research Data IG: https://rd-alliance.org/internal-groups/long-tail-research-data-ig.html

[11] OpenAIRE Guidelines: https://guidelines.openaire.eu

[12] Life Sciences demonstrator: http://ub.unibi.de/oademo

[13] EuropePMC RESTful web service: http://europepmc.org/RestfulWebService

- Citations in other publications
- References to biological databases either manually attributed by data curators or automatically analysed by text-mining
- Medical Subject Headings that are tagged with the publication

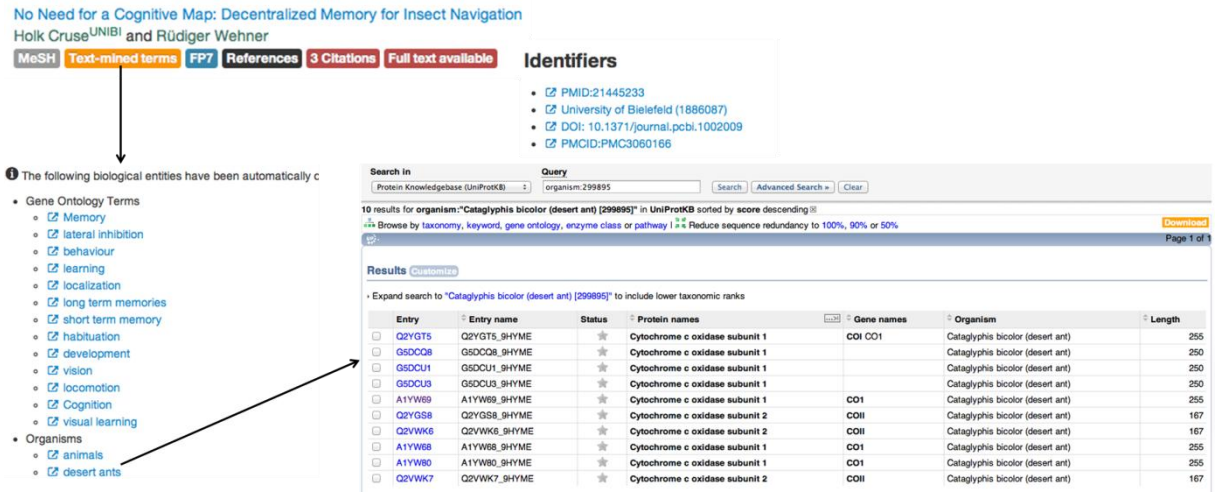

Figure 3 Navigating the information graph in the Life Sciences demonstrator

Social Sciences & Humanities (SSH) demonstrator

The research practices in the SSH domain divides into

- the quantitative social sciences with rather structured methodologies and mainly numerical types of data (surveys),
- the qualitative social sciences and humanities with various methodologies and types of data created, among them interviews, transcripts, digitized material.

The Social Sciences and Humanities demonstrator is derived from the Dutch portal NARCIS[14] that provides research information from CRIS applications and acts as the national aggregator of Dutch institutional repositories. Furthermore, the demonstrator utilizes results from a series of enhanced publication experiments in DRIVER-II (such as

[14] NARCIS: http://www.narcis.nl

using OAI-ORE[15] to describe aggregations of publications with related resources) and continued in tender projects[16] in the Netherlands.

Publications are chosen from two of the tender projects: DataPlus from the social sciences (figure 4) and Veteran Tapes from the humanities. They are enriched with common references to entities:

- with related publications (like citations) in NARCIS,
- with related datasets that are registered in NARCIS,
- with identified authors, editors of the publication and their affiliated organisations,
- with information to the funded projects

and specific references to data entities that depends on the scientific area the publication belongs:

- survey data for DataPlus,
- interview fragments for Veteran Tapes.

The resources and the semantic relations that interlink the resources are encoded as Resource Maps and manually added to the demonstrator. The resource maps are then transformed using XSLT into a human readable web interface.

It was identified that NARCIS contains descriptions of all the entities of the research results. Though the relations between the entities are missing and it is up to data curators or authors to add them manually or to apply sophisticated data mining algorithms.

[15] OAI-ORE: Object Reuse and Exchange - standard to model and describe aggregations of web resources: http://www.openarchives.org/ore/

[16] Enhanced Publication overview at SURF: http://www.surf.nl/en/themas/openonderzoek/verrijktepublicaties/Pages/Default.aspx

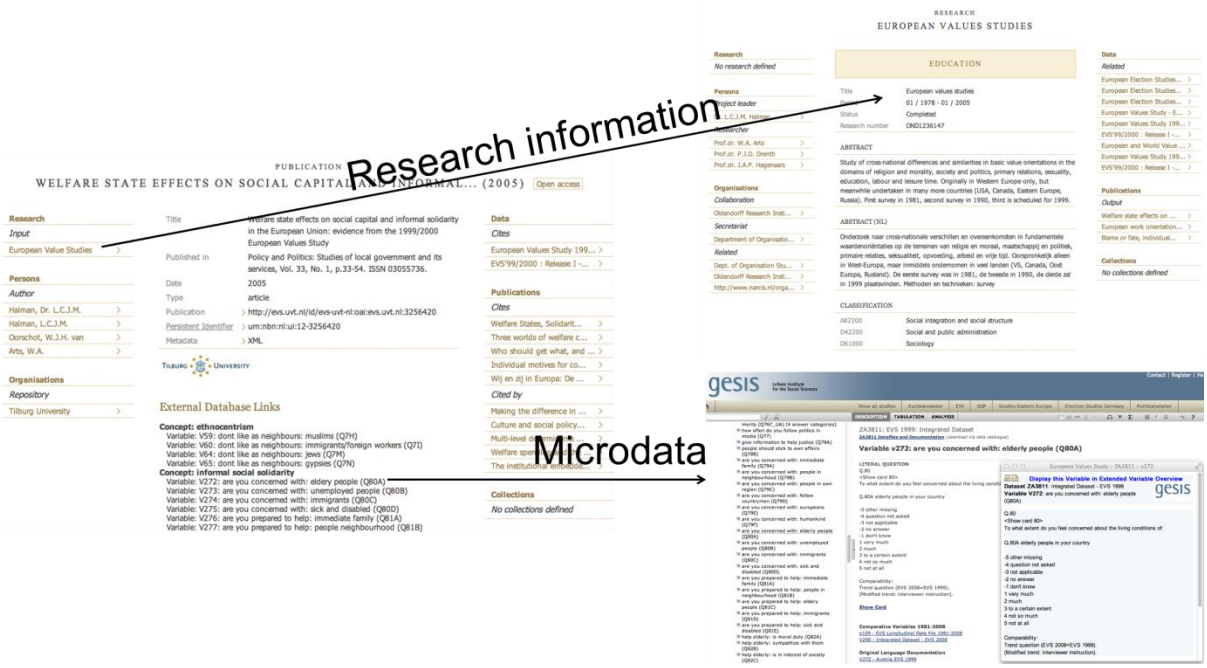

Figure 4 Web presentation of the enriched Social Science example

Both demonstrators show the problem of limiting the references to datasets to a certain number and to select and present only relevant ones. It is often the case that discipline-specific metadata frameworks are bound to discipline and project specific tools and infrastructures and cannot processed and managed in a straightforward manner by generic cross-disciplinary infrastructures. In this case, a user in OpenAIRE should be guided to the subject-specific infrastructure to explore more in-depth findings (such as study concepts and variables, methodologies and research workflows).

3.2 Support of Enhanced Publications in the context of OpenAIREplus

The technical infrastructure builds and extends on the multi-layer scalable software framework D-NET[17]. It provides a rich set of services for aggregation and managing scientific information. Value added functionalities enable users to interact with the OpenAIRE portal as the central gateway to the infrastructure, see figure 5.

OpenAIRE supports various workflows to enter metadata into the information space:

- By collecting (and normalizing) content from data sources (literature and data repositories, CRIS, funder databases, authoritative registries for author and repository identifiers)

[17] D-NET software toolkit: http://d-net.research-infrastructures.eu

- By inferring links from one object (e.g. publication) to other resources and make associations; the approach is based on mining algorithms applied on metadata and fulltext documents

- By claims and feedback about information provided by end users to OpenAIRE about assigning publications to their respective funded project

- Finally by providing links among research results by end users, that can be modelled as aggregations

The OpenAIRE data model was designed to support these various mechanisms, to keep provenance information and assign a level of trust (in descending order from registered data sources to end users). Furthermore it supports the interlinking of all the entities.

A dedicated value-added service – the Enhanced Publication Service – will be implemented that will support Enhanced Publications in either two ways:

1. To present publications in its scientific context by utilizing links in the collected metadata from dataproviders, analysed by inferring algorithms or claimed and entered by end-users.

2. To enable users to interact for constructing and building up Enhanced Publications from Information Space or external resources.

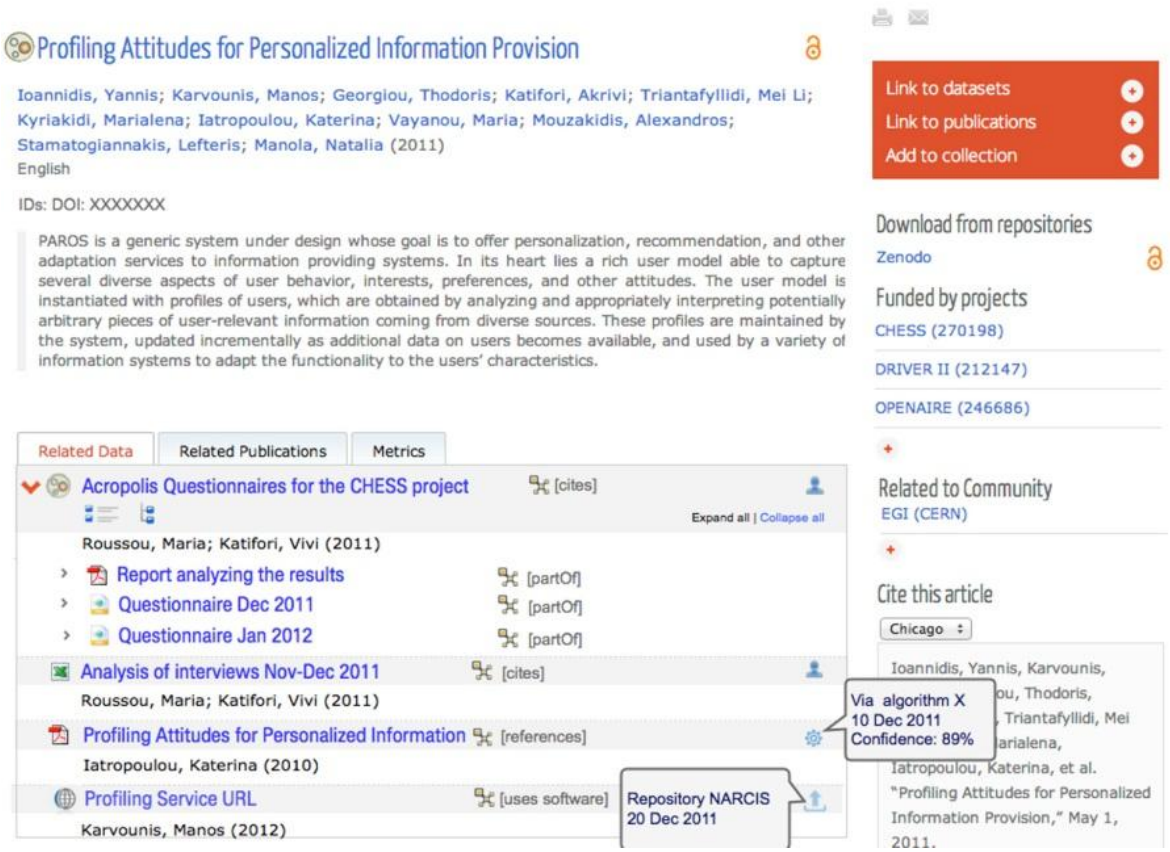

Figure 5 Preview of an enhanced publication in OpenAIRE

CONCLUSIONS

Modern scientific information infrastructures allow scientists to make relations between publications and datasets. PUB at Bielefeld University supports this kind of publication enrichments with links to disciplinary research data services and to other resources in subject-specific information portals.

Conversely scholarly communication infrastructures like OpenAIRE benefit from rich information data sources. Demonstrators have been created and analysed to conceive publication and data management in subject-specific infrastructures. The results can inform the development of OpenAIRE services to manage and link publications and research and to put them in their scientific context.

ACKNOWLEDMENT

This work is partially funded by the European Commission FP7 project OpenAIREplus (FP7-INFRA-2011-2, Grant Agreement no. 283595). It would have not been possible without the support, ideas and experience of partners and people from the OpenAIRE Consortium.

References

- Friedhoff, S., et al.: Social Research Data: Documentation, Management, and Technical Implementation within the SFB 882, SFB 882 Working Paper Series, no.16, Bielefeld: DFG Research Center (SFB) 882 From Heterogeneities to Inequalities. 2013
- Kauppinen, T., Baglatzi, A. and C. Keßler. Linked science: Interconnecting scientific assets, in: Data-Intensive Science,T. Critchlow and K. Kleese-Van Dam, eds, Chapman & Hall/CRC Computational Science, USA, May 2013
- Kinsey, A.: New Europe PMC External Links Service: enabling access to more resources. 2013 http://blog.europepmc.org/2013/07/new-europe-pmc-external-links-service.html
- Lier F, Siepmann F, Paul-Stueve T, Wrede S, Wachsmuth S, Lütkebohle I. Facilitating Research Cooperation through Linking and Sharing of Heterogenous Research Artifacts. In: Sack H, Pellegrini T, eds. Proceedings of the 8th International Conference on Semantic Systems (I-SEMANTICS '12). New York, NY, USA: ACM; 2012: 157–164.

- Hoogerwerf M, Lösch M, Schirrwagen J, et al. Linking Data and Publications: Towards a Cross-Disciplinary Approach. International Journal of Digital Curation. 2013;8(1):244–254.
- IJsbrand Jan Aalbersberg and Judson Dunham and Hylke Koers: Connecting Scientific Articles with Research Data: New Directions in Online Scholarly Publishing. Data Science Journal. 2013;12
- Starr, J.: isCitedBy: A Metadata Scheme for DataCite. D-Lib Magazine. 2011; 17(1/2). doi:10.1045/january2011-starr
- Woutersen-Windhouwer, S. et.al.: Enhanced Publications: Linking Publications and Research Data in Digital Repositories. Amsterdam University Press. 2009 doi: 10.5117/9789089641885
- Wiljes, C. et.al. Towards Linked Research Data: An Institutional Approach. 3rd Workshop on Semantic Publishing (SePublica). CEUR Workshop Proceedings. Aachen; 2013:27-38