Towards Linked Research Data: An Institutional Approach

Cord Wiljes[1], Najko Jahn[3], Florian Lier[2], Thilo Paul-Stueve[2],
Johanna Vompras[3], Christian Pietsch[3], and Philipp Cimiano[1]

[1] AG Semantic Computing, CITEC, Bielefeld University, Germany
{cwiljes,cimiano}@cit-ec.uni-bielefeld.de
http://www.sc.cit-ec.uni-bielefeld.de
[2] Central Lab Facilities, CITEC, Bielefeld University, Germany
{flier,tpaulstu}@cit-ec.uni-bielefeld.de
http://www.cit-ec.de/CLF/CentralLabs
[3] Bielefeld University Library, Bielefeld University, Germany
{najko.jahn,johanna.vompras,christian.pietsch}@uni-bielefeld.de
http://www.ub.uni-bielefeld.de/english/

Abstract. For Open Science to be widely adopted, a strong institutional support for scientists will be essential. Bielefeld University and the associated Center of Excellence Cognitive Interaction Technology (CITEC) have developed a platform that enables researchers to manage their publications and the underlying research data in an easy and efficient way. Following a Linked Data approach we integrate this data into a unified linked data store and interlink it with additional data sources from inside the university and outside sources like DBpedia. Based on the existing platform, a concrete case study from the domain of biology is implemented that releases optical motion tracking data of stick insect locomotion. We investigate the cost and usefulness of such a detailed, domain-specific semantic enrichment in order to evaluate whether this approach might be considered for large-scale deployment.

Keywords: Research Data Management, Scientific Publishing, E-Science, Semantic Web, Ontology, Linked Data.

1 Motivation

The vision of *Open Science* foresees a scientific publishing environment in which research results are made available and shared openly at all stages of the scientific discovery process. Research results in this sense go beyond the traditional publication of a paper and comprise the release of all important and relevant research artefacts including software, analysis scripts, detailed descriptions of experimental conditions etc. As research becomes more and more data-driven, results can only be independently verified and validated if there is full access to underlying data, processing software, experimental protocols, etc. As such, Open Science promises to increase transparency, integrity and efficiency in science, opening up new avenues for scientific discovery, allowing for data to be reused in new contexts and fostering the collaboration across disciplines just to name two of many possible benefits. In fact, making the research process more transparent by making all relevant research data publicly available is regarded more and more as a necessity by the research community itself [1] as well as by funding organisations [2].

The Center of Excellence Cognitive Interaction Technology (CITEC[4]), located at Bielefeld University, is a highly interdisciplinary research institute comprising around 250 researchers from disciplines as varied as computer science, biology, physics, linguistics, psychology and sport science. The goal of CITEC is to conduct basic research in cognition while at the same time producing relevant insights and technology that will provide the basis for a better human-machine interaction. The interdisciplinary nature of CITEC calls for a complex communication across institutional and disciplinary borders. Therefore scientists working at CITEC are generally very open to share their research data.

However, the scientists need support and guidance in releasing their data. Data publication is a complicated procedure that involves many different tasks on various levels: organizational, legal and technical. Surveys have repeatedly shown that scientists want to concentrate on their own research questions instead of bothering with technical questions related to data publication [3]. Therefore, in order to be adopted by scientists, the publication of data has to fulfill the following three conditions [4].

1. *easy*: the publication of data should constitute a minimum effort for the scientist
2. *useful*: data publication should not represent a means in itself but offer an immediate and obvious benefit to the scientific community as well as to the scientist him/herself
3. *citable:* data publications have to be citable so they can be referred to within scientific communication and discourse

The goal of our work is to develop an infrastructure that fulfills these three needs at affordable costs of development and operation. The usefulness of a research data management relies on a successful and flexible integration of heterogeneous data from various sources. We will investigate the role Linked Data can play to solve this challenge.

2 Infrastructure

In the following sections we will describe the individual components of the infrastructure at Bielefeld University and CITEC that enable scientists to publish and manage their scientific output. The ultimate goal is to develop a complete ecosystem of services and solutions necessary on the road towards Open Science.

[4] http://cit-ec.de/

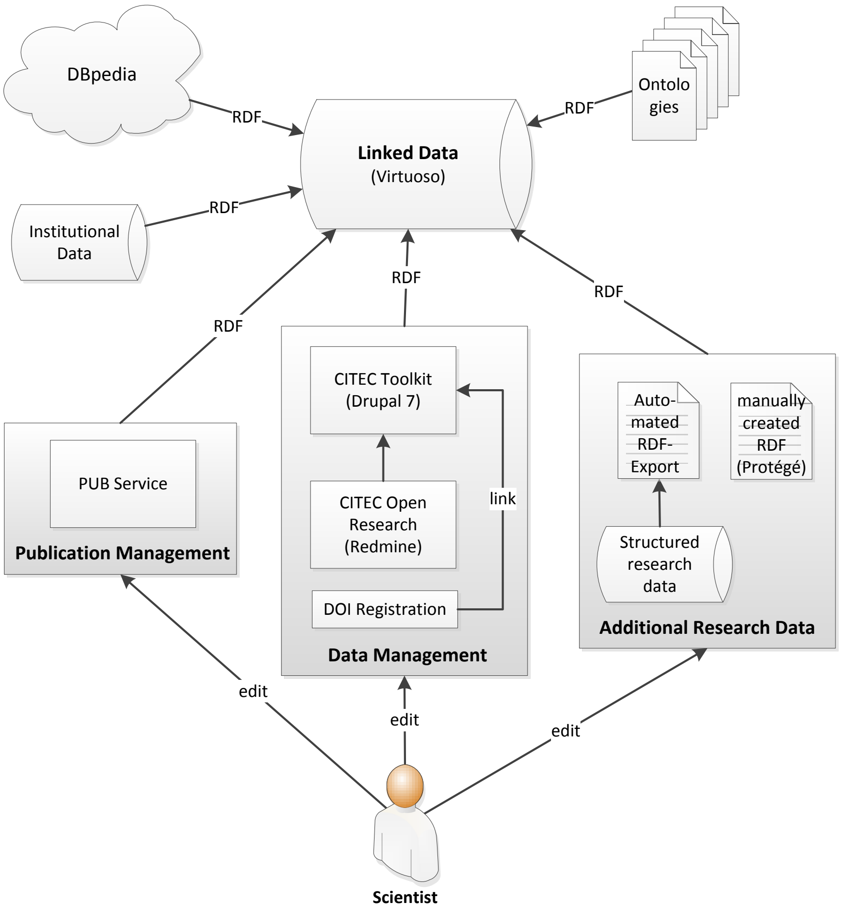

Fig. 1: Open Science infrastructure at Bielefeld University and CITEC with added semantic layer of Linked Data

2.1 Central Services at Bielefeld University

Significant parts of the research infrastructure at Bielefeld University are organised and networked on a university-wide basis.

Publication Management: PUB (Publications at Bielefeld University) provides an overview of publications to which researchers affiliated with Bielefeld University have contributed.[5] This service acts also as an institutional repository allowing to deposit a copy of research publications in accordance with the principle of Open Access to scientific literature. PUB currently has more than 1,000 active users and 35,000 registered publications, of which more than 6,500 provide

[5] http://pub.uni-bielefeld.de

a self-archived Open Access fulltext. 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[5]. Apart from this, PUB exposes its metadata via OAI-PMH harvesting service[6] and SRU search protocoll[7] in various formats (e.g. Dublin Core, MODS).

PUB has been developed under the umbrella of LibreCat[8] – an open source development project of the university libraries of Lund, Ghent and Bielefeld. Here, common toolsets are shared for user-management (Authentication/Authorization), import/export tools (data cleaning tools, import data sources), metadata management (cataloging tools, indexes, lookup lists) and file management (storage, versioning, fixation).

DOI Registration: For research data to become a research asset comparable to publications, it needs to be persistently available and made citeable like publications. This allows to give credit to the researcher and thereby contribute to the scientist's reputation. In addition it makes sure that the data stays available unchanged over time for later verification and re-use. A solution that has become increasingly popular is assigning digital object identifiers (DOIs) for datasets. A DOI uniquely identifies a dataset, can be resolved via a URL and is accompanied by a set of policies to ensure long-term availability and data integrity.

In 2012, Bielefeld University became a DataCite[9] publication agency and since then started an institutional research data registration service that allows scientists to register their research output. Research groups and faculties interested in a DOI registration for their data have to provide the data itself, the metadata based on the required schema, and the URL of a landing page with, if required, an extended version of the metadata.

2.2 Data Management at CITEC

Research data plays a fundamental role in the scientific process, as it is the basis for developing and testing hypotheses. The internet and the availability of cheap storage space has opened the technical basis for large scale data publication. However, it is not yet common practice for scientists to share their research data for other scientists to verify or use in new contexts.

Researchers frequently produce and make use of various research artifacts, e.g., publications, datasets, experiment specifications, software, etc. Often, these artifacts remain underspecified, lacking important information such as which version of a software or dataset was used for a particular publication. Therefore,

[6] http://www.openarchives.org/pmh/

[7] http://www.loc.gov/standards/sru/

[8] http://librecat.org/

[9] http://datacite.org

reproducing experiments and verifying results sometimes becomes unfeasible. To tackle this issue, CITEC has developed a technical platform and procedures for scientists working at CITEC to publish their research data. This platform consists of two interacting components: the *CITEC Open Research* web platform and the *Cognitive Interaction Toolkit.*

CITEC Open Research Platform: The first component, the *CITEC Open Research* web platform[10], is based on the collaborative development environment *Redmine*[11], a Web application for hosting Open Source projects. The CITEC Open Research platform provides scientists with useful project management features like a wiki, a ticketing system or automated notifications by e-mail. In addition, it offers the possibility to upload and publish digital research data onto a central repository, and provides versioning, integrity checking and long-term support. The integrated wiki offers an easy way to provide documentation.

The central idea behind the CITEC Open Research platform is to offer immediate benefits to scientists so that all relevant research data is automatically accumulated during the whole process of a scientific project and can be released at the end of a project with minimal additional effort. As an additional service, at the finishing stage of a project scientists are supported in finding an adequate licensing model for their data.

Cognitive Interaction Toolkit: The second component, the *Cognitive Interaction Toolkit*[12] [6], is focused on integrating and augmenting existing data from the CITEC Open Research platform and is based on the popular content management system *Drupal*[13]. The core concept of the Cognitive Interaction Toolkit comprises of common research artefact types (e.g., publication, open data set, or software), which can be manually created or imported from external sources like the PUB publication repository.

Inside the Toolkit, researchers can enrich the data by adding relations between datasets, software and connected publications to form aggregates of research artefacts. A linked data representation of aggregates and stand-alone entities is created by the platform semi-automatically based on freely definable vocabularies using Drupal's RDF functionalities. By publishing semantically enriched, functionally relevant aggregates on the web, the Cognitive Interaction Toolkit provides a unified view on research artefacts.

Together, the CITEC Open Research web platform and the Cognitive Interaction Toolkit offer a very flexible, low-cost platform for data management. The Web pages with meta-information about versions of research datasets may act as landing pages resolved via DOIs that allow data to become citable.

[10] http://openresearch.cit-ec.de/

[11] http://redmine.org

[12] https://toolkit.cit-ec.uni-bielefeld.de/opendata

[13] http://drupal.org

3 Case Study: Natural Movement Database

We believe that the usefulness of an approach can best be validated by applying it to concrete application cases. Therefore, we implemented a proof of concept scenario to demonstrate the existing infrastructure and to investigate how adding linked data can be beneficial for Open Science. For our case study we chose experiments conducted by the Biological Cybernetics group of CITEC. The motivation and goal of this project has already been presented in detail at SePublica 2012 [7].

Movement is an essential property of animal behaviour. Therefore, understanding movement is an important research question in behavioural neuroscience. The study of movement in biological organisms promises new insights that might be helpful in the creation of artificial systems like robots or embodied agents. At CITEC, movement is being investigated in the context of several research projects. One such project is coordinated by the department of biological cybernetics at Bielefeld University and involves optical motion tracking of stick insects.

The EU project EMICAB[14] conducted at CITEC has set the goal to develop an autonomous hexapod robot [8]. For this, three species of stick insects (*Carausius morosus*, *Aretaon asperrimus*, *Medauroidea extradentata*) are investigated by optical motion tracking. Figure 2 shows a test subject of the species *Aretaon asperrimus* with reflective markers attached. 36 individual test runs of stick insects climbing unrestrained across step obstacles were measured.

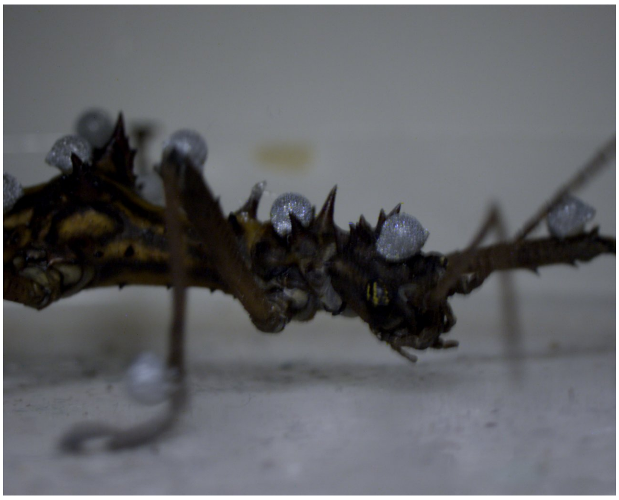

Fig. 2: Stick insect with attached markers (used with permission of Volker Dürr)

[14] `http://emicab.eu/`

As part of this European project, an open-access Natural Movement Database is being constructed. About 4 hours of recording were created and are to be released as open data towards the end of the project EMICAB. The primary trajectory data measured was transferred into joint-angle-files in MATLAB format that use the test subjects' body model and abstracts from the non-reproducible attachment of the markers on the insects body. The metadata about the experiments was transferred from the files via an import script into a SQL-Database whose schema was custom-designed for this purpose. The overall goal is to store the research data in a structured form that allows publication and re-use in future projects. For this purpose, the process of transforming the primary trajectory data into the relational database has been automated. The raw data will be available as downloads under the DOI `http://dx.doi.org/10.4119/unibi/citec.2013.3`.

4 Adding Linked Data

The technical infrastructure at Bielefeld University and CITEC allows the easy publication of research data alongside publications. However, the data uploaded into the repository is highly heterogeneous both in terms of content and format, and requires intensive documentation to become useful to and interpretable by third parties. As the main goal of an institutional repository of research artifacts is to house a variety of potentially very heterogeneous research objects, several questions need to be addressed:

- How can datasets relevant to a research question be successfully retrieved from a large amount of data?
- How can the data remain interpretable over a long period of time, potentially even after the researcher who created it has left the institution?
- How can the platform be open and flexible enough to allow for the addition of yet unforeseen forms of data in the future?
- How can external sources of data be added to the repository? How can the repository's data be exposed to and used by external services?

A promising approach is to add a semantic layer to the data by representing it as linked data. Linked data can be used to build the connections between research data and the publications that are based on it. Because Linked Data is not bound to a fixed schema, it can be extended to fit project-specific needs. By re-using existing ontologies as widely as possible, connections to external datasources are possible.

Institutional Data: To create an ecosystem of linked data from an institutional repository of research artifacts we need to link to other resources inside our university, i.e. to the scientists who created it, to an organization or project it is associated with, to publications it is related to. Therefore, we set up a knowledge base of linked data that contains data about our university, its institutions and researchers. The URI schema for these resources was defined to satisfy the requirements of simplicity, stability and manageability as described in [9] and builds on existing identifiers that had already been used inside the IT infrastructure of Bielefeld University. With the VIVO ontology [10], there exists an ontology that can be readily used and covers most of he basic terms needed to describe entities inside a university. The VIVO ontology builds as much as possible on existing, widely-used vocabularies like FOAF, Dublin Core, BIBO and SKOS. Bielefeld University already had a database of its researchers and their organizational affiliations in a relational database that offers an XML interface. Using Extensible Stylesheet Language (XSLT) this data was transformed into an RDF/XML representation. The same approach was applied on the PUB service via its SRU interface, which exposes metadata as MODS-XML.

Data from Cognitive Interaction Toolkit: The Cognitive Interaction Toolkit is based on the CMS Drupal that allows automated generation of linked data. As a first step the Description of a Project (DOAP) ontology[15] was chosen as a vocabulary. This process provides general metadata at minimal cost, but does not go into the details of the specific research. Our goal is to add additional, research-specific information in the form of linked data while still keeping the overall cost manageable.

Data from DBpedia: In addition to these internal resources, connections to external resources are necessary to explicate the content of the data. To account for the heterogeneity of the data and the fact that the content of future datasets cannot be anticipated, a Linked Data repository is needed that is both commonly accepted and covers a spectrum that is broad enough to contain resources from various disciplines. It has been proposed to use *DBpedia*[16], a Linked Data representation of Wikipedia as a crystallization point for the Web of Linked Data [11]. The meaning of URIs is created in a social process like the meaning of words in natural language [12]. Thus the choice of DBpedia/Wikipedia, the largest collaboratively created collection of human knowledge, as a central hub for the emerging web of linked data seems an obvious one. The English version of the DBpedia knowledge base currently describes 3.77 million things and thereby also covers many topics relevant to science. In addition, DBpedia follows Linked Data principles so it has a human readable version for each URI that explains the URIs meaning and it is very well interlinked to other relevant datasets, forming a central hub in the web of Linked Data. Some relevant URIs for our case study are:

```
http://dbpedia.org/resource/Carausius_morosus
http://de.wikipedia.org/wiki/Kleine_Dornschrecke
http://dbpedia.org/resource/Medauroidea_extradentata
http://dbpedia.org/resource/Optical_motion_tracking
```

[15] `https://github.com/edumbill/doap/wiki`
[16] `http://dbpedia.org`

It is interesting to notice that one of the three species investigated (”Aretaon asperrimus”) has no entry in the english DBpedia, so the German version (”Kleine Dornschrecke”) had to be used.

Ontologies: In addition to data from DBpedia, existing, domain-specific ontologies may be used. For the domain of motion tracking, the Ontology for Shape Acquisition and Processing (SAP) [13] is suitable. However, ontology exploration is rather expensive and requires knowledge about the domain as well as an understanding of ontology design. Therefore the cost of exploring or even extending existing ontologies can only be justified in selected cases, i.e. for very valuable data or in cases where a large amount of data from this domain is to be published.

Research Data of Stick Insect Locomotion: As a proof of concept, we exported the motion tracking data from the relational database containing the stick insect locomotion data as RDF/XML using a PERL-Script. The mapping to appropriate RDF-vocabularies is hard coded in the export script. As the data in the database is very subject-specific and fine-grained, only the most important information for interpreting the data has been exported. Listing 1 presents an excerpt of the RDF code generated by exporting the database.

Listing 1: RDF code describing one motion capture experiment (excerpt)

```
<http://info.cit-ec.de/experiment/1> rdf:type dbpedia:Experiment ,
  rdfs:label "Experiment 1" ;
  dc:date "2010-02-17" ;
  dc:title "Step climbing of stick insect Carausius morosus" ;
  sap:hasAcquisitionConditions
    <http://info.cit-ec.de/AcquisitionCondition/1> ;
  sap:hasAcquisitionDevice <http://info.cit-ec.de/equipment/1> ;
  sr:hasSubject dbpedia:Carausius_morosus ;
  citec:hasExperimentalTechnique dbpedia:Optical_motion_tracking ;
  dc:creator <http://info.uni-bielefeld.de/person/18235412> .
```

The resulting RDF was uploaded into a Virtuoso triplestore and exposed by a SPARQL endpoint. A complete set of DBpedia data from the English and the German edition were also imported into the triplestore. The data can be browsed via Virtuososo Faceted Browser Plug-in and graphically explored via Visual Data Web's *Relfinder*[17] [14]. The data, the SPARQL endpoint and the visualization are available online at `http://motion.linked-open-science.org`.

The SPARQL endpoint allows queries that make use of the additional information contained in the internal and external data. Using DBpedia URIs for the test subjects allows us to connect additional data about these species contained in DBpedia and thereby create advanced retrieval methods for research data. For example, Listing 2 displays a SPARQL query that returns all experiments

[17] `http://www.visualdataweb.org/relfinder.php`

about insects, even though the scientist did not explicitly mention that these species are insects.

Listing 2: SPARQL-query: Give me all experiments that investigate insects!

```
PREFIX citec: <http://cit-ec.de/ontology.owl#>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT ?experiment, ?label
WHERE {
  ?experiment dc:subject ?subject .
  ?experiment rdfs:label ?label .
  ?experiment rdf:type dbpedia:Experiment .
  ?subject a dbo:Insect .
}
```

This example illustrates the usefulness of the Linked Data approach: By combining data from two different sources, questions that were not directly foreseen by the providers of the data can be answered. We expect the power of Linked Data to integrate data from various sources will become more apparent as more and more research datasets are released as linked data.

5 Conclusion and Next Steps

The technical infrastructure at Bielefeld University and CITEC allows scientists to publish research data in a central repository and connect it with the corresponding publications that present the results obtained from the data. By assigning DOIs to the datasets research data becomes citable.

We implemented a case study that demonstrates how this infrastructure can act as a Linked Data hub. Linked Data is used to add a semantic layer that enriches the data with additional internal and external data sources, e.g. by linking to DBpedia as a first step, thereby allowing for a more powerful data retrieval. Publication management has already been widely adopted by scientists at Bielefeld University. Adding the research data has recently been installed and is gathering increasing interest and acceptance. Especially the possibility to obtain DOIs for research data, thereby making it citable, increases the attractiveness of data publication to scientists.

Data publication is still a very support-intensive endeavour: DOI registration requires adhering to service level agreements, legal questions need to be addressed, and adding semantic information as Linked Data is still in its infancy, with domain-specific vocabularies still in formation. The challenge for universities is to create an infrastructure and support for scientists that is affordable, easy to use and presents immediate benefits to the scientists.

As case study we presented the natural movement database that collects motion tracking data of stick insects. Publishing the data as static files using the infrastructure at Bielefeld University has proven to be rather easy and straightforward. However, creating a Linked Data representation of this very domain-specific research data revealed to be rather complex. For future projects, a trade-off between expressiveness and cost needs to be addressed. In general we favour a modular approach that provides for a basic, inexpensive solution but can be enhanced by additional levels of granularity of the semantic enrichment. Which effort is appropriate for the specific data should be decided on a case-by-case basis, depending on the data's nature and value.

We plan to increase the versatility of our system by learning from implementing cases from various disciplines. The goal is to offer a platform that is flexible and powerful enough to adapt to the very heterogeneous requirements from different disciplines, while staying easy to use for the scientists. Our next steps will integrate Linked Data technology more closely with our infrastructure by allowing scientists to directly annotate their data with DBpedia URIs. In addition, we are planning to set up a form-based web front-end that allows scientists to query the linked data stored in the repository's triplestore in an easy and intuitive way.

We believe that successful examples that present a clear benefit to the scientists, both in increasing their scientific reputation and in helping to answer their research questions, will be the best incentive to foster the acceptance of Open Science among scientists.

References

1. Berlin 9 Open Access Conference: Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. http://oa.mpg.de/files/2010/04/berlin_declaration.pdf (2003) [accessed 20-April-2013].
2. Alliance of German Science Organisations: Priority Initiative "Digital Information". http://www.wissenschaftsrat.de/download/archiv/Allianz-digitaleInfo_engl.pdf [accessed 20-April-2013].
3. Feijen, M.: What researchers want - A literature study of researchers' requirements with respect to storage and access to research data. http://www.surffoundation.nl/nl/publicaties/Documents/What_researchers_want.pdf (2011) [accessed 20-April-2013].
4. Smith, V.S.: Data publication: towards a database of everything. BMC research notes **2**(113) (January 2009)
5. Horstmann, W., Jahn, N.: Persönliche Publikationslisten als hochschulweiter Dienst - Eine Bestandsaufnahme. BIBLIOTHEK Forschung und Praxis **34**(2) (2010) 37–45
6. Lier, F., Wrede, S., Siepmann, F., Lütkebohle, I., Paul-Stueve, T., Wachsmuth, S.: Facilitating Research Cooperation through Linking and Sharing of Heterogenous Research Artefacts. In: Proceedings of the 8th International Conference on Semantic Systems. (2012) 157–164
7. Wiljes, C., Cimiano, P.: Linked Data for the Natural Sciences: Two Use Cases in Chemistry and Biology. In: Proceedings of the Workshop on the Semantic Publishing (SePublica 2012). (2012) 48–59

8. Dürr, V., Schmitz, J., Cruse, H.: Behaviour-based modelling of hexapod locomotion: Linking biology and technical application. Arthropod.Struct.Dev. **33**(3) (2004) 237–250
9. Sauermann, L., Cyganiak, R., Völkel, M.: Cool URIs for the semantic web. `http://www.dfki.uni-kl.de/dfkidok/publications/TM/07/01/tm-07-01.pdf` (February 2007) [accessed 20-April-2013].
10. Conlon, M., Corson-Rikert, J.: VIVO: A Semantic Approach to Scholarly Networking and Discovery (Synthesis Lectures on the Semantic Web). Morgan & Claypool Publishers (2012)
11. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - A crystallization point for the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web **7**(3) (September 2009) 154–165
12. Halpin, H.: Social Semantics: The Search for Meaning on the Web (Semantic Web and Beyond). Springer (2012)
13. Papaleo, L., Albertoni, R., Pitikakis, M., Robbiano, F., Vasilakis, G., Hassner, T., Moccozet, L., Saleem, W., Tal, A., Veltkamp, R.: Ontology for Shape Acquisition and Processing 4th Version. `http://www.aimatshape.net/downloads/public/D1-2-2-1-4th-pdf/download` [accessed 20-April-2013].
14. Lohmann, S., Heim, P., Stegemann, T., Ziegler, J.: The relfinder user interface: interactive exploration of relationships between objects of interest. In: Proceedings of the 15th international conference on Intelligent user interfaces. IUI '10, New York, NY, USA, ACM (2010) 421–422

Supplementary Information

The data presented and discussed above, including a SPARQL endpoint that exposes the linked data, is available online at:
`http://movement.linked-open-science.org/linked-data`.

Acknowledgements

We are grateful to Volker Dürr for sharing his research data and helpful insights. This work is funded as part of the Center of Excellence Cognitive Interaction Technology (CITEC) at Bielefeld University.