Linked Data for the Natural Sciences: Two Use Cases in Chemistry and Biology

Cord Wiljes and Philipp Cimiano

Semantic Computing, CITEC, Bielefeld University, Germany
{cwiljes,cimiano}@cit-ec.uni-bielefeld.de
http://sc.cit-ec.uni-bielefeld.de

Abstract. The Web was designed to improve the way people work together. The *Semantic Web* extends the Web with a layer of *Linked Data* that offers new paths for scientific publishing and co-operation. Experimental raw data, released as Linked Data, could be discovered automatically, fostering its reuse and validation by scientists in different contexts and across the boundaries of disciplines. However, the technological barrier for scientists who want to publish and share their research data as Linked Data remains rather high. We present two real-life use cases in the fields of chemistry and biology and outline a general methodology for transforming research data into Linked Data. A key element of our methodology is the role of a *scientific data curator*, who is proficient in Linked Data technologies and works in close co-operation with the scientist.

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

1 Motivation

The World Wide Web was envisioned by its inventor Tim Berners-Lee as a universal information space that enables people to work together and collaborate better [4]. The *Semantic Web* adds an additional layer of *Linked Data* to the Web that allows machines to process the semantics of the data. The Semantic Web has the potential to change the way scientists co-operate and communicate, how they share data, and how they publish their research results. Because science has become more interdisciplinary, the need for the exchange of data between different branches of science has increased dramatically. The Semantic Web offers a solution to this challenge. Data from different fields could be combined in new ways, giving new insights and helping to solve complex problems that require an interdisciplinary approach.

A cornerstone of the scientific method is the requirement that any experiment has to be reproducible [16]. Publishing research data in an open fashion would support for instance:

- the discovery of related datasets, allowing for comparison of results in different contexts, obtained under different experimental conditions etc. This requires that the data is published in some standard format (e.g. RDF) so that Semantic Web search engines can index all data and retrieve and rank all available datasets relevant for a given scientific question or research hypothesis.

- the external validation of data and reproduction of results by other parties. This requires that the data is sufficiently annotated and documented so that the exact experimental conditions can be identified.

Because scientific research data is very valuable, funding agencies have a high interest to prevent duplication of effort and foster the reuse existing data as efficiently as possible [2]. Nowadays, however, primary research data is still mostly stored in closed, non-accessible silos, usually on local hard discs in the scientist's lab. Typically, only the interpreted and aggregated results are made available to the scientific community via standard publication channels (e.g. journal and conference papers).

In general, scientists have been rather reluctant to adopt Semantic Web technologies. The reasons for this reluctance were revealed by several surveys (for a summary cf. [11]). Presumably the most important one is the lack of incentives, i.e. there is so far only limited reward and recognition for publishing research data. In addition, scientists often regard the results of their research as their property and fear others might take unfair advantage of it. Especially in highly competitive research areas this is a major concern.

But there is a growing number of scientists who share the ideal of making research data public and are willing to publicly release their data. These early adopters face another barrier in the form of technical complexity. A considerable effort is necessary to get acquainted with the relevant techniques and paradigms, i.e. Semantic Web and Linked Data technologies. In order to learn more about possible ways to overcome this obstacle, we investigated real-life use cases from natural science departments of our university. We interacted with scientists at our university, and developed a first methodology targeted at lowering the barrier for scientists to release their research data as Linked Data.

Our long-term goal is to develop and validate a methodology with appropriate tools support that facilitates the task of publishing research data as Linked Data as well as to assess and compare the cost, feasibility and ease of use of different approaches systematically. In this paper we describe two use cases we are currently implementing. Using these as a springboard we will explore the promises and possible pitfalls of publishing scientific research as Linked Open Data.

2 Use Cases

The main objective of researchers is to provide answers to open scientific questions in their field, thus advancing their own understanding of key problems and phenomena as well as the one of their research field as a whole. Taking on the additional workload of semantically annotating research data will only be considered if it does not put too much strain on their time budget and can be integrated with their research work. To overcome this obstacle we decided to investigate an approach of co-operation and support. We contacted scientists who are willing to share their data and offered to take care of the technical side of the publication of research data as Linked Data while the scientists contribute their domain knowledge.

We selected two current research projects carried out by natural science departments from Bielefeld University: one from chemistry and one from biology. Both topics are highly interdisciplinary and produce research data that is potentially relevant to researchers from other disciplines. Both scientists were open to the idea of sharing their research data and were willing to contribute their domain knowledge. In the following we will present these two use cases as well as the involved scientists in more detail.

2.1 Chemistry: Glass Transition of Atmospheric Aerosols

Thomas Koop is a professor of Physical Chemistry at Bielefeld University (Germany). He is co-founder and executive editor of the open access journal *Atmospheric Chemistry and Physics*[1]. His research interests include the properties of atmospheric aerosols and their influence on cloud formation. In September 2011, he published a paper on the glass transition of organic aerosols [15].

Aerosols, which consist of floating particles in the air, are an important factor in many atmospheric processes, like light scattering and cloud formation. According to new insights, water soluble organics can form amorphous solids (glasses) in the upper troposphere (i.e. at 8-15 km height), which inhibit ice crystal formation, thereby affecting cirrus cloud formation [18].

In order to quantify the magnitude of this effect, data about the *glass transition temperature* T_g of various substances known to be present in the atmosphere is needed. Because glass transition temperatures are not collected in chemical databases, Thomas Koop conducted an extensive, manual literature research, which took about 100 hours of work. He collected the resulting 596 T_g values from 22 publications in a large spreadsheet-table and supplemented them by additional information like provenance, measurement methods and additional comments.

The corresponding publication [15] does not publish the full list but results aggregated from this data in the form of digrams (an example is shown in Figure 1). A publication of the full dataset as Linked Data, enriched by a semantic representation of the supplementary data, could be very helpful for other scientists and prevent duplication of effort.

Use Case: As a use case we take the example of a researcher in chemistry who wants to collect glass transition temperatures of aerosols. Instead of compiling the data manually from published research articles as Thomas Koop did, our scientist would use Semantic Web search engines to collect relevant data and use

[1] `http://atmos-chem-phys.net/`

appropriate SPARQL queries to aggregate results as needed. Issues that need to be paid attention to are provenance, data quality as well as the fact that different vocabularies might have been used in publishing the data, so that vocabulary harmonization is a crucial part of the process.
Some sample competency questions a researcher could pose to the dataset are:

Q: Give me all glass transition temperatures of organic compounds!

Q: Give me all glass transition temperatures of amino acids measured by differential scanning calorimetry!

Q: Which substances form glasses at temperature and pressure conditions in the troposphere?

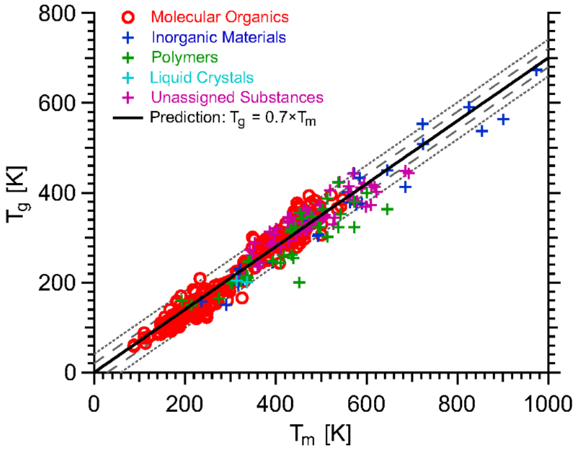

Fig. 1: Graph of the evaluated dataset of glass transition temperatures plotted against melting temperatures. (From [15] - Reproduced with permission of The Royal Society of Chemistry)

2.2 Biology: Natural Movement of Stick Insects

Volker Dürr is a professor of Biological Cybernetics at Bielefeld University (Germany). His research interests include the question of how insects adapt their locomotion behaviour to the context of the situation. He coordinates the EU project EMICAB[2], which has the objective to develop an autonomous hexapod robot.

Insects like the stick insect (Figure 2) can walk on rough terrain, climb obstacles, and use their legs for other behavioural tasks such as searching or reaching [7]. These complex movements are coordinated by a fairly small, experimentally amenable and reasonably well-studied nervous system [5]. Because of the

[2] http://emicab.eu/

resource-efficient information processing for solving complex behavioural tasks, the analysis and modelling of insect locomotion have been proposed as a basis for improving artificial autonomous walking robots [10].

The movement of stick insects can be measured by marker-based motion capturing: markers are attached to the body of the insect and tracked by an infrared camera system. The resulting trajectories (time-ordered *xyz*-coordinates) describe the movement of the insect in space. Volker Dürr's group recorded several hours of locomotion sequences from different stick insect species by motion capture. The interpretation of the trajectory data is dependent on the body morphology and the position of the markers on the body. Motion capture datasets have been released in the past, but without specifying the anatomy of the test subject and the exact marker locations, such that these datasets are of limited use outside their original purpose.

A novel approach is to provide sufficient annotation for calculating joint angle time courses for all degrees of freedom from the trajectory data. This would allow the data to be interpreted and reused in other contexts. Pioneering this approach, the EU project EMICAB will make such calculated data publicly available alongside the experimental raw data and metadata about the experimental conditions under which the data was obtained. A semantic annotation of these datasets would greatly improve their retrieval and interpretation by potential future users.

Use Case: A researcher interested in insect motion might download this dataset and extract or recompute the joint angle time courses for all degrees of freedom, thus being able to simulate the organism or compare it to his own results for other or the same organism. The challenge is to incorporate enough information in the data about how the joint angle time courses have been computed so that the comparison is meaningful.
Competency questions the dataset has to answer include:

```
Q: Give me all motion capture datasets about insects!
Q: How large is the complete dataset?
Q: What data is necessary to reproduce the experiment?
```

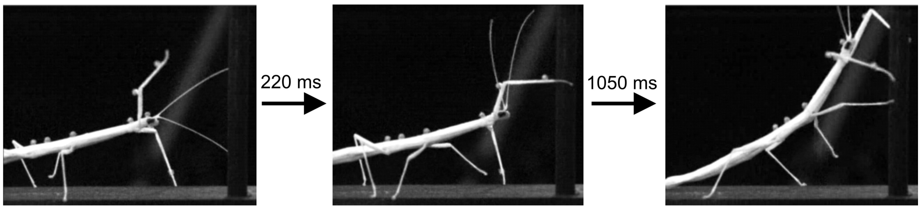

Fig. 2: Stick insect movement with markers for motion capture attached. (Reproduced with permission of Volker Dürr)

3 Methodology

A key element in our methodology is the role of the *scientific data curator*. The data curator's task is to translate the methods and results of scientific research into Linked Data. His role could be compared to that of an investigative reporter: he is not an expert in the domain he is describing, but he is proficient at finding out what is essential and relevant. He asks the scientist the right questions to find out what others need to know to understand and reuse the data. Further, he should be proficient in Semantic Web and Linked Data technologies.

Transferring scientific research into Linked Data can be viewed as a project which requires a joint effort between the scientist and the data curator, such that a close co-operation and constant feedback is essential during all phases of the project. We propose a methodology which involves seven consecutive tasks:

Task 1: Kick-off Meeting
The kick-off meeting is the first meeting of the scientist and the data curator and marks the start of the project. In the kick-off meeting the project members get to know each other and lay the groundwork for the future co-operation. The data curator interviews the scientist about his research interests and gives an introduction into the technology of Linked Data. Ideas and expectations are exchanged in order to build a common understanding of the goal and scope of the project, which will be defined in the next step.

Task 2: Goal Definition
Following the kick-off meeting, the data curator formulates a proposal for the goal and the scope of the project and subsequently refines it by feedback from the scientist. As a main tool at this stage of the project we formulate competency questions, which can be used as tests to make sure that the data contains all relevant information, and to choose vocabularies to represent the data.

For our use cases the goal is to capture all *relevant* data, i.e. the experimental results and all information necessary to reproduce these results. A more lightweight approach could concentrate only on the data *essential* for interpreting the experimental results. The most comprehensive scope would be to use all *available* data, even the pieces that seem irrelevant for the reproducibility of the experiment - but could prove relevant in the future or in other contexts.

Task 3: Knowledge Acquisition
The data curator acquires domain specific knowledge. He achieves this by interviews with the scientists and reading the papers which are based on the experiments. His aim is not to become an expert himself but to get an overview and basic understanding in a short period of time. In addition he collects data which might already be available in structured or semi-structured form.

The glass transition temperatures were collected in a large spreadsheet table with informal comments and undocumented color-coding. The stick insect movement was available in a relational database.

Task 4: Ontology + LOD Exploration

At this stage, the data curator explores existing vocabularies and ontologies that could be reused. He has to thoroughly investigate them in order to evaluate their applicability and usefulness for his task. In addition he is looking for existing Linked Open Data (LOD) datasets that can be linked to. Interlinking and reuse is extremely important, because most of the usefulness of the data lies in its connection to external data. If concepts or resources are involved for which no existing vocabularies or datasets can be located, the data curator will create them. The understanding, evaluation, disambiguation and alignment of existing ontologies is the most important and labour intensive task of the whole process because many existing ontologies are not properly documented.

For our use case in chemistry, several ontologies for the domain of chemistry exist, e.g. CHEMINF [13], ChemAxiom [1] or ChEBI [9]. They differ substantially in scope and complexity. For our use case in biology the *Shape Acquisition and Processing* (SAP) ontology [8] is relevant, which covers the domain of movement data, forms, and virtual characters. The first dataset to look for possible links is *DBpedia*[3], which offers a wealth of concepts and is well dereferenceable for human readers.

Task 5: Implementation

If one or more ontologies have been selected, the data is encoded using the technological tool most appropriate. This could range from the mapping of an existing database, using annotation software or even manual encoding.
Figure 3 presents sample RDF code for encoding a glass transition temperature.

```
:PinicAcid a :ChemicalSubstance ,
  :hasCASNumber "[473-73-4]" ;
  :hasName "Pinic Acid"@en ;
  :hasProperty
    [ a :GlassFormationPoint ;
      dc:source "http://dx.doi.org/10.1039/C1CP22617G" ;
      :hasValue "268.1"^^xsd:float ;
      :hasUnit :Kelvin ;
      :hasStandardDeviation "4.8"^^xsd:float ;
      :hasMeasurementCondition
        [ a :MeasurementPressure ;
          :hasValue "101325"^^xsd:float ;
          :hasUnit :Pascal
        ] ;
      :hasExperimentalTechnique :differentialScanningCalorimtery
    ] .
```

Fig. 3: RDF-representation of the glass transition temperature of pinic acid.

[3] http://dbpedia.org

Task 6: Publication

The data is published, either by uploading the code representing the knowledge to a web server or by importing it into a triplestore. This task essentially completes the project. A SPARQL endpoint should be also provided ideally so that the data can be queried flexibly as needed.

Task 7: Monitoring

Subsequently, the usage of the published data is continuously monitored, for example by looking at SPARQL queries generated by third parties. This can help to improve and refine the data selection and implementation.

Parallel Task: Documentation

Documenting is not a separate task but is done parallel to the other tasks. Like in all projects, documentation plays an important role in forming a common understanding between project members, to proceed from one task to the next, and to enable others to understand and continue the work in the future.

The individual tasks are not strictly linear but feedback loops to earlier tasks are possible if a subsequent task should require correction or refinements to an earlier task. Figure 4 shows an overview of the proposed methodology and possible feedback loops between the tasks.

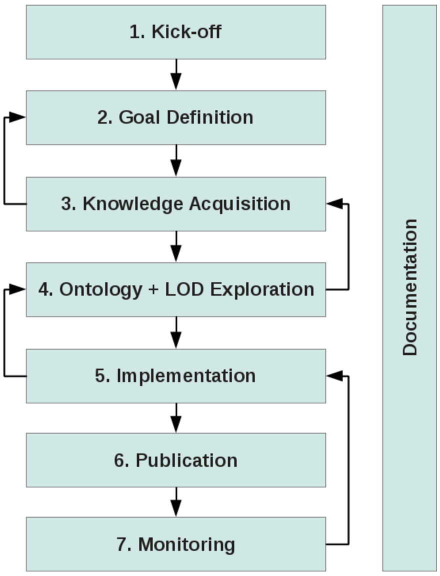

Fig. 4: Workflow for the semantification of research data.

4 Related Work

The *Open Science* movement aims to make scientific publications and research data publicly available. Numerous initiatives have formed over the last few years that put these ideals into practice. Open access journals create alternatives to the old publication system, e.g. *Atmospheric Chemistry and Physics*[4], which has been using an open review system for 10 years and has the highest impact factor of all 68 journals in the field of meteorology and atmospheric sciences. Even traditional publishers are beginning to embrace the new technologies, like Elsevier did with its *Grand Challenge*[5]. Universities are setting up public repositories of research data, e.g. VIVO[6] or *Potsdam Mind Research Repository*[7], which gives access to peer-reviewed publications and additional data and scripts for analyses and figures. Large Datasets have been opened, like the *Human Genome Project*[8] or the *Sloan Digital Sky Survey*[9]. The W3C's *Health Care and Life Sciences Interest Group* (HCLSIG)[10] created a knowledge base of RDF data from the domains of health care and the life sciences. Social networks like myExperiment[11] allow scientists to publish and share their scientific workflows. With all of these the ideas of Open Science are gradually changing the way scientific research is done.

One of the main tasks of our methodology is the elicitation of knowledge from the domain experts. Methodologies for knowledge extraction have a long tradition in knowledge management (cf. [14]). Especially relevant to our approach is the work on ontology engineering [17]. A data curator does not have the primary goal of creating an ontology or vocabulary, but he may find it necessary to develop a vocabulary, or extend an existing one, if no existing ontology for a specific task can be found. In any case, he needs a good understanding of methodologies for the creation and evolution of ontologies, in order to evaluate and apply them.

For an efficient creation of Linked Data several approaches have been developed, either by automated translation or by tool-support for the author. Four kinds of approaches can be distinguished:

1. Export from existing sets of structured data: relational databases, like the ChEBI database[12], which collects data about chemical substances, are exported into Linked Data by mapping database fields to a vocabulary. The D2R Project[13] exposes the content of a relational database as Linked Data.

[4] http://atmos-chem-phys.net/
[5] http://www.elseviergrandchallenge.com/
[6] http://vivoweb.org/
[7] http://read.psych.uni-potsdam.de/pmr2/
[8] http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
[9] http://www.sdss.org/
[10] http://www.w3.org/wiki/HCLSIG
[11] http://www.myexperiment.org/
[12] http://www.ebi.ac.uk/chebi/
[13] http://d2rq.org

2. Export from content management systems (CMS): Drupal[14], WordPress[15] can publish editorial content as Linked Data using pre-selected vocabularies.
3. Automated extraction of data from scientific publications by text-mining techniques: Several methods for automatic extraction of bibliographic metadata have been developed [12]. The OSCAR3 [6] programme identifies chemical terms by natural language processing.
4. Semantic annotation of papers either by editors or by scientists: within the *Prospect* project[16] for instance, the Royal Society of Chemistry (RSC) has taken the approach to have papers semantically enriched not by the scientist but by editors (cf. [3] for an overview).

Which of these approaches is the best one for a specific task depends on the goal and the scope of the individual project. Because our goal is to develop deeper insights into how existing vocabularies and ontologies can be reused in the process of publishing Linked Data, we have decided to carefully evaluate and select the most appropriate vocabularies instead of converting the data to RDF using some automatic approach (e.g. RDB2RDF[17]). In the future we plan to compare the results with those of more automatic approaches.

5 Conclusion and Future Work

We presented two use cases from chemistry and biology that we are currently working on. The specific aim of these projects is to publish the relevant scientific research data as Linked Data, i.e. the results of the experiments and the experimental set-ups necessary to reproduce the results. We proposed a methodology that is characterized by a close co-operation between a scientist and a scientific data curator, who translates the scientists' domain knowledge into Linked Data. This preliminary methodology will be validated and refined empirically as the implementation progresses.

In preparing the projects we found that all scientists we interacted with are interested in the ideas and possibilities of Linked Open Data. But only few of them are willing to contribute and invest their data, time and knowledge. A close co-operation between scientist and data curator is highly important for the success of the project. Therefore trust is essential. The scientist must be sure that his data is handled responsibly and that his wishes regarding its publication are respected.

So far we have defined the goals and the scope of both projects and elicited the relevant domain knowledge. We are currently in the process of evaluating suitable existing ontologies and Linked Data from other datasets we could link to. Because our goal to publish all relevant research data is rather ambitious,

[14] `http://drupal.org/`

[15] `http://wordpress.org`

[16] `http://www.rsc.org/Publishing/Journals/ProjectProspect/`

[17] `http://www.w3.org/2001/sw/rdb2rdf/`

the cost involved with each of the steps is high. Especially the exploration and evaluation of existing ontologies has proven to be complex and time consuming.

Our long-term objective is to contribute to the formation of an open research infrastructure by empowering scientists to publish their research as Linked Data. Towards this goal, appropriate methodologies for the transformation of research data into Linked Data are needed. In combination with shared ontologies and tool support, we expect these to be the foundation for scientists to adopt the new technology of Linked Data. Our hypothesis is that the role of a *scientific data curator* as proposed in this paper is a key function towards facilitating this development.

After completing the two use cases we will perform a thorough analysis of the resulting datasets and of the overall process. Focus will be put on the question of the cost involved for each of the tasks. Our next step will be the development of criteria for choosing, combining and expanding existing ontologies. In future work we plan to use our manually created Linked Data as a gold standard for the evaluation of less expensive, semi-automatic or fully automatic solutions.

Acknowledgements

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

References

1. Adams, N., Cannon, E., Murray-Rust, P.: Chemaxiom - an ontological framework for chemistry in science. Available from Nature Precedings: `http://dx.doi.org/10.1038/npre.2009.3714.1` (2009)
2. Alliance of German Science Organisations: Priority Initiative "Digital Information". Retrieved April 12, 2012, from: `http://www.wissenschaftsrat.de/download/archiv/Allianz-digitale%20Info_engl.pdf` (June 11, 2008)
3. Attwood, T.K., Kell, D.B., McDermott, P., Marsh, J., Pettifer, S.R., Thorne, D.: Calling international rescue: knowledge lost in literature and data landslide! Biochemical Journal **424**(3) (2009) 317–333
4. Berners-Lee, T., Fensel, D., Hendler, J.A., Lieberman, H., Wahlster, W., eds.: Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. MIT Press, Cambridge, MA (2005)
5. Büschges, A., Akay, T., Gabriel, J.P., Schmidt, J.: Organizing network action for locomotion: Insights from studying insect walking. Brain Res. Rev. **57**(1) (January 2008) 162–171
6. Corbett, P., Murray-Rust, P. In: High-throughput identification of chemistry in life science texts. Volume 4216. Springer Berlin Heidelberg (2006) 107–118
7. Cruse, H., Dürr, V., Schilling, M., Schmitz, J.: Principles of insect locomotion. In Arena, P., Patanè, L., eds.: Spatial temporal patterns for action-oriented perception in roving robots. Springer, Berlin (2009) 43–96

8. De Floriani, L., Hui, A., Papaleo, L., Huang, M., Hendler, J.: A semantic web environment for digital shapes understanding. In: Proceedings of the semantic and digital media technologies 2nd international conference on Semantic Multimedia. SAMT'07, Berlin, Heidelberg, Springer-Verlag (2007) 226–239
9. Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., Ashburner, M.: Chebi: a database and ontology for chemical entities of biological interest. Nucleic Acids Research **36**(suppl 1) (2008) D344–D350
10. Dürr, V., Schmitz, J., Cruse, H.: Behaviour-based modelling of hexapod locomotion: linking biology and technical application. Arthropod Struct Dev **33**(3) (2004) 237–50
11. Feijen, M.: What researchers want - a literature study of researchers' requirements with respect to storage and access to research data. Retrieved April 12, 2012, from SURFfoundation: `http://www.surffoundation.nl/nl/publicaties/Documents/What_researchers_want.pdf` (February 2011)
12. Groza, T., Grimnes, G., Handschuh, S., Decker, S.: From raw publications to linked data. Knowledge and Information Systems 1–21
13. Hastings, J., Chepelev, L., Willighagen, E., Adams, N., Steinbeck, C., Dumontier, M.: The chemical information ontology: Provenance and disambiguation for chemical data on the biological semantic web. PLoS ONE **6**(10), DOI 10.1371/journal.pone.0025513: `http://dx.doi.org/10.1371%2Fjournal.pone.0025513` (10 2011)
14. Holsapple, C., ed.: Handbook on knowledge management. International handbooks on information systems. Springer, Berlin (2003)
15. Koop, T., Bookhold, J., Shiraiwa, M., Pöschl, U.: Glass transition and phase state of organic compounds: dependency on molecular properties and implications for secondary organic aerosols in the atmosphere. Phys. Chem. Chem. Phys. **13** (2011) 19238–19255
16. Popper, K.R.: The Logic of Scientific Discovery. Hutchinson, London (1959)
17. Sure, Y., Staab, S., Studer, R.: Ontology Engineering Methodology Handbook on Ontologies. In Staab, S., Studer, R., eds.: Handbook on Ontologies. International Handbooks on Information Systems. Springer, Berlin, Heidelberg (2009) 135–152
18. Zobrist, B., Marcolli, C., Pedernera, D.A., Koop, T.: Do atmospheric aerosols form glasses? Atmos. Chem. Phys. **8**(17) (2008) 5221–5244