Semantic Technologies for Consolidating Structured Data and Unstructured Documents in Biodiversity Research

Birgitta KÖNIG-RIES[1] and Udo HAHN[2]

[1]Heinz-Nixdorf-Endowed Chair for Distributed Information Systems
Friedrich-Schiller-Universität Jena, Germany
E-Mail: birgitta.koenig-ries@uni-jena.de

[2]Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Germany
E-Mail: udo.hahn@uni-jena.de

Abstract

Over the last few years, the large-scale generation and content-focused access to data has increasingly become a driving force for scientific progress in many scientific disciplines including the life and earth sciences. Flexible information management and the support for advanced data analysis thus turn out to be keys to productive and innovative research activities. In this position paper, we argue that there is not a single semantic tool or technique that will offer the necessary support. What is needed is rather a suite of semantic tools with focus on the seamless interoperability of data. Such a framework could, in the end, provide novel opportunities for carrying out science *in silico* rather than, only, *in situ*.

1 Introduction

Over the last few years, the large-scale generation and content-focused access to data has increasingly become a driving force for scientific progress in many scientific disciplines including the life and earth sciences. Throughout this paper, we will use biodiversity research as an example discipline. However, we believe, that the requirements and solutions we point out are not restricted to this special discipline but apply equally to many other data-heavy subject areas.

As an example, consider the key questions that the *German Center for Integrative Biodiversity Research, iDiv*,[1] aims to address:

i. How can we detect and quantify biodiversity?

ii. How does biodiversity emerge and how is it maintained?

iii. What are the consequences of biodiversity for the functioning of ecosystems?

[1] www.idiv-biodiversity.de

iv. How can we effectively safeguard biodiversity?

Quite obviously, *the* basis for answering such questions are integrated and seamlessly synthesizable data. For biodiversity research these data typically include a wide range of heterogeneous data such as taxonomic data (e.g., species lists), observation and measurement data, different types of sensor data, remote sensing data, sequence data, geographic information and so on.

Flexible information management and the support for advanced data analysis thus turn out to be keys to productive and innovative research activities. In this position paper, we argue that there is not a single semantic utility or technique that will offer the necessary support. What is needed is rather a suite of semantic tools with focus on the seamless interoperability of data. Such a framework could, in the end, provide novel opportunities for carrying out science *in silico* rather than, only, *in situ*. In other words, combining hitherto isolated data and making them interoperable might allow entirely novel forms of computational hypothesis generation and testing, thus complementing and interacting with *in situ* experimentation in the lab or field.

Particular challenges to achieve this goal arise from the fact that scientific data today appear, from a computational point of view, in two fundamental shapes. In a *structured* form they populate scientific databases, simple file structures or experimentalists' spreadsheets, in an *unstructured* form they are dispersed in scientific publications, reports and memos. We thus have to make structured and unstructured data semantically homogeneous so that their diverse origins (though fully traceable) and heterogeneous formats are no longer obstacles to common use.

At the core of data homogenization lie procedures for automatically uncovering their implicit semantics, a task that is already hard for structured data but even harder for unstructured data. However, the sheer mass of already available data collections and the exponential growth rates of structured and unstructured data tend to overwhelm experimental scientists in their daily work. Under these circumstances, solely providing homogeneous access to tons of data by way of data integration and interoperability is not enough for breaking the data complexity barrier. In order to raise the level of ‘data awareness’ of all scientists and avoid the duplication of efforts due to lacking access to hidden, yet existing knowledge, we rather foresee the need for *semantic data abstraction* and based on it an entirely new breed of scientific information infrastructure. These emerging systems will be (pro)active knowledge exploration workbenches rather than merely passive data containers with attached retrieval front-ends.

Accordingly, the ultimate goal is the provision of a semantics-driven engine that will not only allow the proper management of all kinds of (un)structured data collected or produced by individual data providers but, furthermore, will transparently interlink these data to support scientific reasoning at the desktop, i.e. an *in silico* hypothesis generation, testing and adaptation system based on a focused and coherent view of unified and interoperable scientific data, their (hypothetical) interactions and causal relationships that can be inferred from them.

Based on our previous work on biodiversity data management and biological text mining and ongoing work within the CRC Aquadiva[2], in this position paper, we propose the necessary steps to achieve this goal. More concretely, we will first outline a scenario of semantics-driven scientific data and knowledge management (Section 2), then review existing work both by others and our own (Section 3) and finally elaborate on near-term and longer-term steps that need to be taken to ensure the best possible tool support for researchers based on these premises (Section 4).

2 A Scenario for Semantic Data Integration and Interoperability with Focus on Biodiversity Research

Currently, the majority of structured data resides in physically separated local or publicly available data bases and, very common in the life sciences, lab-internal spreadsheets. Almost all of these data have their own data formats, both at the technical storage and the logical data organization level. Accordingly, data exchange and data integration are hard to achieve and proceed, if at all, between selected single instances of such repositories based on data structure-specific manual mappings at both schema and data level.

On the other hand, unstructured data contained in publications are, for now, accessible by Google-style document retrieval systems at a shallow keyword-based level only. These systems find possibly relevant documents (with respect to a problem-specific query) but deeper content can usually not be targeted on a larger scale. As a consequence, many relevant data contained in publications are virtually not available for scientists, except from those papers they (or database curators) have read on their own and, possibly, transferred into some database or content management format.

Hence, we envisage the following requirements for consolidating structured data and unstructured publications for biodiversity research:

- **Data management.** The basic requirement is the provision of a *common logical and technical platform* for storage of and access to the data produced by different scientific groups/labs. Such a platform can build on existing software as discussed in Section 3.
- **Data integration.** Heterogeneous collections of structured data which sit in public and lab-internal data silos need to be made *uniformly and thus homogeneously accessible* via a common scientific data management framework. This unification could be achieved by *assigning semantic metadata* descriptions to data sets, which formally reflect the underlying meaning of those raw data.
- **Data interoperability.** To achieve data interoperability, i.e., the ability to seamlessly combine structured and unstructured data stemming from different sources, structured, semantically integrated data from databases and unstructured data from scientific publications will need to become part of a novel, hitherto non-existing semantically homogeneous information management framework. A particular focus needs to be on text analytics pipelines in order to structure the information

[2] http://www.aquadiva.uni-jena.de/

contained in unstructured publications and alto so provide an appropriate semantic integration level. Their format corresponds to the metadata descriptions for structured data and, thus, eases a semantically transparent integration of both structured and (originally) unstructured verbalized data.

- **Semantic computing over data.** The final goal is to supply biodiversity researchers with a semantically rooted scientific problem solving platform which allows extensive quantitative (statistics, simulations, etc.) and qualitative (formal reasoning) computations over structured quantitative and qualitative data. This will allow researchers to derive novel facts and new evidence from the underlying data that also take into account the data's uncertainty, credibility, consistency, etc.

3 State of the Art

Data play a more and more crucial role for scientific progress (HEY et.al 2009). Data synthesis and integration are key to answering the big scientific questions in many sciences, in particular multi-disciplinary ones such as ecology. This requires the provision of powerful data management platforms REICHMAN et.al. (2011). Consequently, over the last decade, a plethora of such systems have been developed ranging from loosely coupled networks to strongly integrated databases, from generic to domain or even project-specific solutions, and from mere repositories to sophisticated workbenches. Still, fully adequate solutions for data management in ecology and related disciplines are missing, up until now (BENDIX et.al. 2012). Remaining challenges include the need to balance the flexibility necessary to accommodate a wide variety of heterogeneous data from different disciplines with ease of data synthesis, the seamless integration of workflow support, handling of uncertainty and provenance and easy integration of external structured and unstructured data sources. Individual systems address some of these issues, but no comprehensive, fully integrated solution addressing all of them already exists.

Scientific workbenches not specifically tailored to the domain at hand, e.g. Kepler (LUDÄSCHER et.al. 2006) and Taverna (MISSIER et.al. 2010), offer seamless integration of different tools and workflow modelling capabilities, but do not yet provide sophisticated integration of heterogeneous data and, in particular, of unstructured data. Similarly, VisTrails (CALLAHAN et.al. 2006) provides the linkage of raw data with simulation data and allows for interactive visualizations to be embedded in publications. It thus provides integration of structured in unstructured data, but does not support extraction of information from text. On the other hand, none of the systems specifically geared towards ecological research, such as DiversityWorkbench[3], BEFData[4], FOR816dw and BExIS (LOTZ et.al. 2012), support the full range of desiderata listed above.

For geographic information systems, the potential of semantic integration has long been acknowledged (see, e.g., FONSECA & CÂMARA, G. (2002). Using ontologies for integrated geographic information systems, numerous works have been undertaken, e.g., (KOUBARKIS, M. et al. 2012.), in general, though, uptake in practice is still slow.

[3] http://diversityworkbench.net

[4] http://china.befdata.biow.uni-leipzig.de/

The probabilistic databases community (SUCIU et.al. 2011) offers techniques for provenance and uncertainty management as well as data analysis support, e.g. in projects such as Trio (AGGARWAL 2009), Panda (IKEDA & WIDOM 2010), Mystiq (RÉ 2009) and Hazy[5]. However, these techniques have not yet been fully integrated in scientific data management platforms. A first glimpse of what might become possible with such integrated designs can be traced in the GeoDeepDive subproject of Hazy[6]. Here, for one sample database and a clearly delimited domain, the envisioned integration of structured and unstructured data has already been successfully realized. In order to be useful for more diverse domains and for heterogeneous data sources, the approach still lacks flexibility and scalability.

Given the growing amount not only of structured data, but also of unstructured, textually encoded data, the need to support scientists in extracting relevant information from scientific texts has been recognized for biodiversity science and crucial requirements for that domain have been identified (THESSEN et.al. 2012). Basically, two kinds of text mining applications are in focus for semantic document analysis. First, the recognition of so-called named entities, i.e. objects of interest in a certain science domain, such as proteins/genes, species or tissues in the biological domain, and the extraction of semantic relations holding between named entities. As an example, consider the sentence "IL-7 influences FOXP3+CD4+ regulatory T cells" which contains "IL-7" and "FOXP3+CD4+ regulatory T cells" as named entities of the type Protein and T-cell, respectively, as well as the semantic relation "Influence (IL-7 , FOXP3+CD4+ regulatory T cells)" (for surveys, cf. NADEAU & SEKINE (2007) and ZHOU et.al. (2014), respectively). The latter is already close to structured formats such as RDF which can easily be computationally interpreted as structured data. Since the domains covered by biodiversity research are large and diverse, existing text mining devices (mainly developed for the biology, medicine and chemistry domain) have to be complemented by ones from currently uncovered domains (geology, ecology, geophysics, bio-physics, bio-chemistry, marine science, forestry and agriculture research, etc.) (see, e.g. current efforts, such as the Poseidon system for the extraction of marine microbiology-related information (RADOM et.al. 2012, THESSEN & PARR 2014)). Reasonable new entity types might be Geologica (with mentions of earth surface, subsurface, critical zone, well, rock, soil, sediment, sandstone, forest, grassland, farmland, etc.), Physica (with mentions of atmosphere, energy, fluid dynamics, gas, etc.) or Meteorologica (with mentions of climate, water, rain, snow, sunshine, etc.), while entirely novel relation types specific for biodiversity research have to be specified, as well. The resulting classifiers will complement the collections of already existing classifiers, e.g. for genes/proteins, species and chemicals. In our own work, we build, e.g., on text mining pipelines which include top-performing named entity recognizers (WERMTER et.al. 2009) and relation extractors (BUYKO et.al. 2011).

As mentioned above, biodiversity research encompasses a wide range of disciplines. For basically all of the domains involved, efforts are underway to formalize the domain knowledge in terms of standardized vocabularies or even ontologies (for a survey, cf. MADIN et.al. (2008)). Quite obviously, such formalization is needed as the fundament for all integration efforts (see, e.g. the Owl-based specification of the Oboe ontology (MADIN

[5] http://research.cs.wisc.edu/hazy

[6] http://hazy.cs.wisc.edu/hazy/geo/

et.al. 2007) or efforts directed at an Ecological Metadata Language (LEINFELDER et.al. 2010)). Only if a shared understanding of the meaning of the terminology of the fields involved can be achieved, will it become possible to automatically compute relationships between data, to detect commonalities or conflicts among data from different sources, and ultimately, to computationally create and test hypotheses *in silico* (MYERS & ATKINSON 2013)

Examples for such efforts in the biodiversity domain include (but are certainly not limited to) ChEBI[7] for chemistry, ENVO for biological environments (BUTTIGIEG et.al. 2013), BCO for biological collections (WALLS ET.AL. 2014), SWEET for environmental science (RASKIN & PAN 2005), NAL Agricultural Thesaurus[8] for agriculture, Biocomplexity Thesaurus[9] for geoscience and ecology, OBOE[10] for scientific observations, INSPEC Thesaurus[11] for physics or the taxonomies, standards and vocabularies developed by the core biodiversity community (e.g. Catalogue of Life[12], ABCD[13], or International Plant Name Index[14]). Some of these knowledge repositories are hosted by biomedical ontology portals such as Obo[15] and BioPortal.[16] Since many of these ontologies have overlaps, efforts have to be made to align terminologies and ontologies properly using good practice protocols (BEISSWANGER & HAHN 2012).

4 Next Steps

To reach the goals described above, a number of steps can be identified that have to be taken immediately, based on the state of the art outlined in the previous section. Once these tasks have been performed, more ambitious and long(er)-term steps can be addressed.

Let us first take a look at what can—and should—be done immediately within any larger biodiversity research project and that we currently undertake in the context of CRC AquaDiva.

First of all, any such project needs a platform for scientific data management which ensures that data produced within this project can be stored in one central place and safeguarded for future use. This, of course, includes the annotation of data with appropriate metadata, quality control of data, etc. Such frameworks can be based on the existing platforms mentioned in the previous section. Additionally, the projects should adopt data policies governing shared (and ultimately public) access to data and ensure that data management, its importance and how to do it, are properly taught to young scientists.

[7] http://www.ebi.ac.uk/chebi/
[8] http://agclass.nal.usda.gov/
[9] http://www.usgs.gov/core science systems/csas/biocomplexity thesaurus/
[10] https://semtools.ecoinformatics.org/oboe
[11] http://www.theiet.org/resources/inspec/about/ records/ithesaurus.cfm
[12] http://www.catalogueoflife.org/
[13] http://wiki.tdwg.org/ABCD
[14] http://www.ipni.org/lsids.html
[15] http://www.obofoundry.org/
[16] http://bioportal.bioontology.org

Second, the existing platforms should be extended to address the needs for data integration described above. Which extensions are necessary obviously depends on which platform you look at. For example, for platforms like BExIS (LOTZ et.al. 2012) or DiversityWorkbench currently used in German biodiversity research projects, this would mean to add mechanisms for mapping concepts used by the scientists to ontological entities and to exploit this semantic knowledge for advanced personalized semantic search facilities, similar to the ones we developed in (BAKALOV et.al 2012) thus helping the researchers fully exploit the available data. We are currently developing such an ontological mapping using mostly the CHEBI and SWEET ontologies and are integrating the respective component into the BExIS platform. Though these processes are currently mainly carried out manually, we expect to support this mapping by (semi)automatic means (KNOBLOCK et.al. 2012).

These first two steps are meant to provide experimental scientists with 'stable' working environments, i.e., pay tribute to their long-standing tradition in individually tailored, yet typically isolated forms of data management. Among the structured formats, we encounter different varieties of proprietary in-house databases as well as lab-internal spreadsheet collections. If collaborative projects possess powerful common information infrastructures which ease reorganizing data in a semantically coherent, transparent way, a lot is gained already. Even more added value can be achieved by, third, including seamless access to structured data provided by public databases and networks. Such integration will only succeed, though, if the efforts put into vocabulary and ontology development are maintained and efforts towards bridging ontologies from neighboring domains are strengthened.

Less obvious at first sight, though equally important is, fourth, content-based access to data contained in unstructured textual formats such as scientific publications. Comparable with database and spreadsheet resources, there are multiple technical encodings of texts (e.g., different types of PDFs) that have to be unified prior to feeding them to content-oriented text analytics. Since many scientific publications nowadays contain data collections and assorted reference material from the experiments, which is discussed in the textual body of a publication, format diversity at the plain data level is further increased in terms of semi-structured data (data files with tabs, XML encodings, etc.).

The main clue to harmonizing these heterogeneous resources, (semi-)structured data repositories and unstructured documents, in terms of transparent integration and interoperability, i.e., to successfully performing at least Steps 2 to 4 above, will be to locate and make explicit the meaning of these data through a battery of semantic technologies (HAHN 2013). These include, among others, methods for data and knowledge reformatting (e.g. using RDF/S), knowledge rendering and querying (e.g. using OWL and SPARQL dialects, respectively), formal ontologies to represent the underlying domain knowledge, as well as comprehensive ways of knowledge exploration (e.g. using interactive visualization).

Finally, all further steps depend on high (or at least known) data quality. The assessment of the quality of experimental data, among many other things, depends on experimental conditions, instrumentation and tooling, sampling and statistical methods being employed, approved reproduceability, as well as data ageing effects. In addition, the origin of the data (database provider, type of journal, author(s) of a publication, etc.) is of great importance. The quality of data expressed in verbal form is further dependent on the capabilities (error rate) of text analytics systems but also explicit linguistic markers intrinsic to the verbalization of scientific knowledge and findings in publications (for instance, due to the use of

lexical markers such as “there is evidence for X”, “we might stipulate that X holds”, “X is highly likely”, etc. indicating that ‘X’ is not a plain undisputed fact but uncertain or hypothetical in nature (HAHN& ENGELMANN 2014). It is therefore essential to account for different degrees of evidence, certainty, and thus credibility of data, irrespective of whether they come from a structured (database) or an unstructured (textual) source (see, e.g. NUTTLE et.al. 2009)). This is true not only for primary data directly stored in or retrieved from a data source, but also for derived data obtained by integrating data from different sources and their provenance.

Once these five issues have been successfully tackled, more visionary tasks can be addressed: Their common goal is to ‘boost’ the uniform knowledge resources (already at the ontology level of logical specification) by the exploitation of various sorts of inference modes—classical deductive ones but also non-standard ones such as explicitly hypothetical abductive reasoning (FLACH & HADJIANTONIS 2000). Additionally, probabilistic reasoning (PEARL 1988, KOLLER & FRIEDMANN 2009) will play a role here in order to incorporate considerations of data quality and credibility at the computational level. As of now, it is not entirely clear, which approaches (or combinations of approaches) are the most promising here. Clearly, more research is needed.

Finally, once these first reasoning problems have been resolved, causal modelling and the incorporation of the above-mentioned non-standard inference techniques into a model of scientific, i.e., rational reasoning and scientific hypothesizing seems to be possible [(PEARL 2000). From a system’s perspective, this means that it will become possible to realize hypothesis exploration systems which, step by step, will mirror the progress of different levels of semantically modelling scientific reasoning processes. We envisage here a workbench for simulating alternative assumption scenarios that will allow *in-silico* finding, testing and adapting of hypotheses. In order to render empirical evidence to these *in-silico* derived claims, validations experiments always have to be run. We claim that this perspective opens a whole new avenue of doing experimental and data-heavy science.

5 Summary and Conclusions

In this paper, we have argued that semantic technologies offer the chance for revolutionary new ways to perform computationally grounded science in biodiversity research. We have discussed how the seamless integration of *a priori* structured and unstructured data will allow to exploit existing knowledge to its full potential and to create new *knowledge in silico*. We summarized already existing building blocks towards such a solution and outlined where additional research is needed to make this vision become a reality for biodiversity researchers.

References

COENEN, R. (2000), Konzeptionelle Aspekte der Entwicklung von Nachhaltigkeitsindikatoren. In: TA-Datenbank-Nachrichten, 2, 47-53.

FABY, H. & KOCH, W. G (2006), Medienrevolution, Kommunikation und Kartographie: Interdependenzen zwischen dem Wandel gesellschaftlicher Systeme der Kartenherstel-

lung und der Kartennutzung. In: Wiener Schriften zur Geographie und Kartographie, 17. Wien.

AGGARWAL, C. (2009), Trio: A system for data uncertainty and lineage. In: Managing and Mining Uncertain Data, pages 1–35.

BAKALOV, F., MEURS, M., KÖNIG-RIES,B., SATELI, B.,WITTE, R., BUTLER, G. & TSANG, A. (2012). Personalized semantic assistance for the curation of biochemical literature. In: Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on, pages 1–4. IEEE.

BEISSWANGER, E. & HAHN, U. (2012), Towards valid and reusable reference alignments: ten basic quality checks for ontology alignments and their application to three different reference data sets. In: Journal of Biomedical Semantics, 3((Suppl 1)):S4.

BENDIX, J., NIESCHULZE, J. & MICHENER W. (2012), Data platforms in integrative biodiversity research. In: Ecological Informatics, 11:1–4.

BUTTIGIEG, P., MORRISON, N., SMITH, B., MUNGALL, C., & LEWIS, S. & ENVO CONSORTIUM, THE (2013). The Environment Ontology: contextualising biological and biomedical entities. in: Journal of Biomedical Semantics, 4, 43.

BUYKO, E., FAESSLER, E., WERMTER, J. & HAHN, U.(2011), Syntactic simplification and semantic enrichment: trimming dependency graphs for event extraction. In: Computational Intelligence, 27(4):610–64.

CALLAHAN, S., FREIRE, J., SANTOS, E., SCHEIDEGGER, C., SILVA, C. & AND VO H. (2006), Vistrails: visualization meets data management. In: Proceedings of the 2006 ACM SIGMOD international conference on Management of data, pages 745–747. ACM.

FONSECA, F., EGENHOFER, M., AGOURIS, P. (2002). Using ontologies for integrated geographic information systems. In: Transactions in GIS, *6*(3), 231-257.

FLACH, P. & HADJIANTONIS, A. (EDS.) (2000), ABDUCTIVE AND INDUCTIVE REASONING. SPRINGER, 2000.

HAHN, U. (2013) , Semantic Technologies: A Computational Paradigm for Making Sense of Qualitative Meaning Structures, in: B.-O. Küppers, U. Hahn & S. Artmann (Eds.), Evolution of Semantic Systems, pages 151-173, Springer, 2013

HAHN, U., & ENGELMANN, C. (2014). Grounding epistemic modality in speakers' judgments. in: Trends in Artificial Intelligence. PRICAI 2014 – Proceedings of the 13th Pacific Rim International Conference on Artificial Intelligence. Gold Coast, Australia, 1-5 Dec, 2014, 654–667.

HEY, A., TANSLEY, S. & AND TOLLE, K., (2009), The fourth paradigm: data-intensive scientific discovery. Microsoft Research Redmond, WA,

IKEDA, R. & WIDOM, J. (2010), Panda: A system for provenance and data. In: IEEE Data Eng. Bull., 33(3):42–49.

KOUBARAKIS, M., KARPATHIOTAKIS, M., KYZIRAKOS, K., NIKOLAOU, C., VASSOS, S., GARBIS, G., ... & GREGOR, R. (2012). *Building virtual earth observatories using ontologies and linked geospatial data* (pp. 229-233). Springer Berlin Heidelberg.

KNOBLOCK, C., SZEKELY, P., AMBITE, J., GOEL, A., GUPTA, S., LERMAN, K., MALLICK, P., MUSLEA, M. & TAHERIYAN, M. (2012). Semi-automatically mapping structured sources into the Semantic Web. in: ESWC 2012 - Proceedings of the 9th Extended Semantic Web Conference. Crete, Greece, 2012, 375-390.

KOLLER, D., & FRIEDMAN, N. (2009). Probabilistic Graphical Models. Cambridge/MA: MIT Press.

LEINFELDER, B., TAO, J., COSTA, D., JONES, M., SERVILLA, M., O'BRIEN, M. & AND BURT, C., (2010),. A metadata-driven approach to loading and querying heterogeneous

scientific data. Ecological Informatics, 5(1):3–8.

LOTZ, T., NIESCHULZE, J., BENDIX, J., DOBBERMANN, M. & KÖNIG-RIES, B., (2012). Diverse or uniform? Intercomparison of two major german project databases for interdisciplinary collaborative functional biodiversity research. In: Ecological Informatics.

LUDÄSCHER, B., ALTINTAS, I., BERKLEY, C., HIGGINS, D., JAEGER, E., JONES, M., LEE, E., TAO, J. & ZHAO, Y., (2006). Scientific workflow management and the kepler system. In: Concurrency and Computation: Practice and Experience, 18(10):1039–1065.

MADIN,J., BOWERS, S., SCHILDHAUER, M., KRIVOV, S., PENNINGTON, D. & AND VILLA, F., (2007). An ontology for describing and synthesizing ecological observation data. In: Ecological Informatics, 2(3):279–296.

MADIN, J., BOWERS, S., SCHILDHAUER, M. & JONES, M. (2008), Advancing ecological research with ontologies. In: Trends in Ecology & Evolution, 23(3):159–168.

MYERS, TRINA, & ATKINSON, IAN (2013). ECO-INFORMATICS MODELLING VIA SEMANTIC INFERENCE. IN: INFORMATION SYSTEMS, 38, 16-32.

MISSIER, P., SOILAND-REYES, S., OWEN, S., TAN, W., NENADIC, A., DUNLOP, I., WILLIAMS, A., OINN, T. &, GOBLE, C. (2010), Taverna, reloaded. In: M. Gertz, T. Hey, and B. Ludaescher, editors, SSDBM 2010, Heidelberg, Germany.

NADEAU, D. & AND SEKINE, S. (2007),. A survey of named entity recognition and classification. In: Lingvisticae Investigationes, 30(1):3–26.

NUTTLE, T., BREDEWEG, B., SALLES, P. & AND NEUMANN, M. (2009), Representing and managing uncertainty in qualitative ecological models. In: Ecological Informatics, 4(5-6):358–366.

PEARL, JUDEA (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo/CA: Morgan Kaufmann.

PEARL, JUDEA (2000). Causality: Models, Reasoning, and Inference. Cambridge, U.K.: Cambridge University Press.

RADOM, M., RYBARCZYK, A., KOTTMANN, R., FORMANOWICZ, P., SZACHNIUK, M., GLÖCKNER, F., REBHOLZ-SCHUHMANN, D. & BLAZEWICZ, J. (2012), Poseidon: An information retrieval and extraction system for metagenomic marine science. In: Ecological Informatics, 12((Nov)):10–15.

RÉ, C. (2009),. Managing large-scale probabilistic databases. PhD thesis, University of Washington.

REICHMAN, O., JONES, M. & SCHILDHAUER, M. (2011),. Challenges and opportunities of open data in ecology. In: Science(Washington), 331(6018):703–705.

SUCIU, D., OLTEANU, D., RÉ, C. & KOCH, C. (2011), Probabilistic databases. In: Synthesis Lectures on Data Management, 3(2):1–180.

THESSEN, A., CUI, H. & AND MOZZHERIN, D. (2012), Applications of natural language processing in biodiversity science. In: Advances in Bioinformatics, doi:10.1155/2012/391574.

THESSEN, ANNE E., & PARR, CYNTHIA SIMS (2014). KNOWLEDGE EXTRACTION AND SEMANTIC ANNOTATION OF TEXT FROM THE ENCYCLOPEDIA OF LIFE. IN: PLOS ONE, 9, E89550.

WALLS, R., DECK, J., GURALNICK, R., BASKAUF, S., BEAMAN, R., BLUM, S., BOWERS, S., BUTTIGIEG, P., DAVIES, N., ENDRESEN, D., GANDOLFO, M., HANNER, R., JANNING, A., KRISHTALKA, L., MATSUNAGA, A., MIDFORD, P., MORRISON, N. Ó TUAMA, É., SCHILDHAUER, M., SMITH, B., STUCKY, B., THOMER, A., WIECZOREK, J., WHITACRE, J. & WOOLEY, J. (2014). Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. in: PLoS

ONE, 9, e89606.
WERMTER, J., TOMANEK, K. & HAHN, U. (2009), High-performance gene name normalization with GeNo. Bioinformatics, 25(6):815–821.
ZHOU, D., ZHONG, D. & HE, Y. (2014). Biomedical relation extraction: from binary to complex. in: Computational and Mathematical Methods in Medicine, 298473.