Bulgarian Academy of Sciences

Institute of Information and Communication Technologies
Department of Linguistic Modelling and Knowledge Processing

Pensoft Publishers

Dissertation

submitted towards the degree

"Ph. D."

in doctoral program "Informatics" (01.01.2012)
professional field 4.6 "Informatics and Computer Science"

The Open Biodiversity Knowledge Management System in Scholarly Publishing

---

Author:
M. Sci. Viktor Senderov

Academic adviser:
Prof. Dr. Lyubomir Penev
Academic consultant:
Assoc. Prof. Dr. Kiril Simov

Sofia, 2019 г.

"*The semantic web is the future of the internet and always will be.*"

Peter Norvig, Director of Research at Google

BULGARIAN ACADEMY OF SCIENCES
PENSOFT PUBLISHERS

Abstract

Institute of Information and Communication Technologies

"Ph. D."

The Open Biodiversity Knowledge Management System in Scholarly Publishing

by Viktor Senderov

OpenBiodiv is a publicly-accessible production-stage semantic system running on top of a reasonably-sized biodiversity knowledge graph. It stores biodiversity data in a semantic interlinked format and offers facilities for working with it. It is a dynamic system that continuously updates its database as new biodiversity information becomes available by connecting to several international biodiversity data publishers. It also allows its users to ask complex queries via SPARQL and a simplified semantic search interface.

Contents

Availability of data and software

openbiodiv.net	OpenBiodiv portal (Beta), supported by Pensoft
graph.openbiodiv.net	OpenBiodiv SPARQL endpoint and LOD dataset
github.com/vsenderov/dissertation-v2	Dissertation LaTeX source code
github.com/pensoft/openbiodiv-o	OpenBiodiv ontology source code
github.com/pensoft/rdf4r	R library for working with RDF
github.com/pensoft/ropenbio	XML converters used with RDF4R
github.com/pensoft/OpenBiodiv	OpenBiodiv documentation

Introduction

Importance of the topic

The desire for an integrated information system serving the needs of the biodiversity community dates at least as far back as 1985 when the Taxonomy Database Working Group (TDWG)—later renamed to Biodiversity Informatics Standards but retaining the abbreviation TDWG—was established[1]. In 1999, the Global Biodiversity Information Facility (GBIF) was created after the Organization for Economic Cooperation and Development (OECD) had arrived at the conclusion that *"an international mechanism is needed to make biodiversity data and information accessible worldwide"* (*What is GBIF?*). The Bouchout declaration (*Bouchout Declaration* 2014) crowned the results of the European Union–funded project *pro-iBiosphere* that lasted from 2012 to 2014 and was dedicated to the task of creating an integrated biodiversity information system. The Bouchout declaration proposes to make scholarly biodiversity knowledge freely available as Linked Open Data (LOD). A parallel process in the U.S.A. started even earlier with the establishment of the Global Names Architecture, GNA (Patterson et al., 2010; Pyle, 2016b).

In 2014, the Horizon 2020 BIG4 consortium was formed between academia and industry dedicated to advancing biodiversity science. The project's mission statement reads *"BIG4—Biosystematics, Informatics and Genetics of the big 4 insect groups: training tomorrow's researchers and entrepreneurs"* (University of Copenhagen et al., 2014). An important member of the consortium is the academic publishing house and software company, Pensoft Publishers. It publishes several dozen well-known open access taxonomic journals[2] and, as a signatory of the Bouchout declaration, was a prime candidate to push the vision for an Open Biodiversity Knowledge Management System (OBKMS) forward. The presented Ph.D. project is based at Pensoft Publishers and at the Institute of Information and Communication Technology (IICT) of the Bulgarian Academy of Sciences with the goal to follow through pro-iBiosphere's vision.

Previous work

Due to the interdisciplinary nature of the thesis, this section will focus on two areas: (a) knowledge bases and Linked Open Data and (b) biodiversity publishing.

[1] A web page with the history of TDWG dating back to 1985 can be viewed under `http://old.tdwg.org/past-meetings/`; however, a lot of the links are unfortunately broken and the page needs some maintenance.

[2] For example, ZooKeys, PhytoKeys, MycoKeys, and Biodiversity Data Journal (BDJ).

Knowledge bases and Linked Open Data

We shall start by first introducing *knowledge bases* and *knowledge-based systems*. We use the two terms interchangeably but tend to write the longer variant, knowledge-based system, when we want to emphasize aspects of the knowledge base that are not related to the underlying facts store (database).

It is useful to form one's concept of knowledge-based systems both by looking at explicit definitions and by looking at several examples of knowledge bases in practice. The term was already being widely discussed by the 1980's (Jarke et al., 1989) and early nineties (Harris et al., 1993) and was understood to mean the utilization of ideas from both database management systems (DBMS) and artificial intelligence (AI) to create a type of computer system called *knowledge base management system* (KBMS). Harris et al., 1993 writes that the characteristics of a knowledge base management system are that it contains *"pre-stored rules and facts from which useful inferences and conclusions may be drawn by an inference engine."* We should note that the phrase "pre-stored rules" comes from the time of first-generation AI systems that were rule-based. Recently, there has been progress in incorporating statistical techniques into databases (Mansinghka et al., 2015); however, in this project we are working with the classical rule-based definition. In other words, a knowledge base is, in our understanding, *a suitable database tightly integrated with a logic layer.*

Another relatively recent development in knowledge-based systems has been the application of the Linked Data principles (Heath and Bizer, 2011). In fact, most existing knowledge bases emphasize the community aspects of making data more interconnected and reusable. Examples include Freebase (Bollacker et al., 2008), which was recently incorporated in WikiData (Vrandečić and Krötzsch, 2014; Pellissier Tanon et al., 2016), DBPedia (Auer et al., 2007), as well as Wolfram|Alpha (*Wolfram|Alpha, Making the wolrd's knowledge computable*) and the Google Knowledge Graph (Singhal, 2012). What these systems have in common is that an emphasis is placed not only on the logic layer allowing inference but on a unified information space: these systems act as nexus integrating information from multiple places and they follow to various degrees the principles of Linked Open Data (LOD).

Linked Open Data (Heath and Bizer, 2011) is a concept of the Semantic Web (Berners-Lee et al., 2001), which, when applied properly, ensures that data published on the Web is reusable, discoverable, and most importantly ensures that pieces of data published by different entities can work together. We will discuss the Linked Data principles and their application to OpenBiodiv in detail in Chapter 3.

Leveraging these developments, modern knowledge bases place a bigger emphasis on interlinking data rather than on developing a complex inference machinery. There has been critique of the idea of bundling logic in the database layer as such bundling leads to increased complexity (Barrasa, 2017). The critique can be summarized with two points. First, bundling the logic near the data (especially when it is excessive for the task at hand) can lead to drastic performance decreases[3]. Second, the developing of new techniques (e.g. machine learning) can make the existing deep logic layer obsolete. Our view is that data is the commodity which is much more valuable, and the inference strategy (be it a rule-based logic layer, or a statistical machine learning technique) can be replaced as computational science moves forward. These ideas lead to an interesting conundrum in the choice of a database technology discussed in the subsequent sections.

[3] We will compare the performance of the stronger Web Ontology Language (OWL) logic layer with a weaker RDF Schema (RDFS) logic layer in Chapter 3. Resource Description Framework (RDF) is a data model for storing statements about things discussed later.

Finally, a knowledge-based system ultimately needs to include user-interface components (UI's) and application programming interfaces (API's) or an application layer. These serve as the point-of-contact for human-computer, or computer-computer interaction, and are crucial to the success of any such system.

Biodiversity publishing

In the biomedical domain there are well-established efforts to extract information and discover knowledge from literature (e.g. Rebholz-Schuhmann et al., 2005; Momtchev et al., 2009; Williams et al., 2012). The biodiversity domain, and in particular biological systematics and taxonomy (from here on in this thesis referred to as *taxonomy*), is also moving in the direction of semantization of its research outputs (Agosti, 2006; Patterson et al., 2006; Kennedy et al., 2005; Penev et al., 2010a; Tzitzikas et al., 2013). The publishing domain has been modeled through the Semantic Publishing and Referencing Ontologies, SPAR Ontologies (Peroni, 2014). The SPAR Ontologies are a collection of ontologies incorporating, amongst others, FaBiO, the FRBR-aligned Bibliographic Ontology (Peroni and Shotton, 2012), and DoCO, the Document Component Ontology (Constantin et al., 2016). The SPAR Ontologies provide a set of classes and properties for the description of general-purpose journal articles, their components, and related publishing resources. Taxonomic articles and their components, on the other hand, have been modeled through the TaxPub XML Document Type Definition (DTD)—also referred to loosely as XML schema—and the Treatment Ontologies (Catapano, 2010). While TaxPub is the XML-schema of taxonomic publishing for several important taxonomic journals (e.g. ZooKeys, PhytoKeys, Biodiversity Data Journal), the Treatment Ontologies are still in development and have served as a conceptual template for OpenBiodiv-O (discussed in Chapter 2).

Taxonomic nomenclature is a discipline with a very long tradition. It transitioned to its modern form with the publication of the Linnaean System (Linnaeus, 1758). Already by the beginning of the last century, there were hundreds of taxonomic terms in usage (Witteveen, 2015). At present the naming of organismal groups is governed by the International Code of Zoological Nomenclature, ICZN (International Commission on Zoological Nomenclature, 1999) and by the International Code of Nomenclature for algae, fungi, and plants, Melbourne Code (*International code of nomenclature for algae, fungi and plants (Melbourne code)* 2012). Due to their complexity (e.g. ICZN has 18 chapters and 3 appendices), it proved challenging to create a top-down ontology of biological nomenclature. Example attempts include the relatively complete NOMEN ontology (Dmitriev and Yoder, 2017) and the somewhat less complete Taxonomic Nomenclatural Status Terms, TNSS[4].

There are several projects that are aimed at modeling the broader biodiversity domain conceptually. Darwin Semantic Web, Darwin-SW (Baskauf and Webb, 2016) adapts the previously existing Darwin Core (DwC) terms (Wieczorek et al., 2012) as RDF (RDF Working Group, 2014). These models deal primarily with organismal occurrence data.

Modeling and formalization of the strictly taxonomic domain has been discussed by Berendsohn (Berendsohn, 1995) and later, e.g., in (Franz and Peet, 2009; Sterner and Franz, 2017). Noteworthy efforts are the XML-based Taxonomic Concept Transfer Schema (Taxonomic Names and Concepts Interest Group, 2006) and a now defunct

[4]Even though it is unknown to the authors whether TNSS was published in peer-reviewed literature, remnants of it can still be found on GitHub, e.g. under `https://github.com/pensoft/OpenBiodiv/blob/master/ontology/contrib/taxonomic_nomenclatural_status_terms.owl`.

Taxon Concept ontology. Very recently, the TDWG community has attempted to resurrect the Taxon Concept ontology with the Taxonomic Names and Concepts Interest Group. The group discussions can be accessed under `https://github.com/tdwg/tnc`. Interestingly the very first GitHub issue discussed OpenBiodiv-O and the possibility of its adoption as a TDWG standard.

By the time the OpenBiodiv project started in June 2015, a number of articles had been previously published on the topics of linking data and sharing identifiers in the biodiversity knowledge space (Page, 2008), unifying phylogenetic knowledge (Parr et al., 2012), taxonomic names and their relation to the Semantic Web (Page, 2006; Patterson et al., 2010), and aggregating and tagging biodiversity research (Mindell et al., 2011). Some partial discussion of OBKMS was to be found in the science blog iPhylo (Page, 2014, 2015). The legal aspects of the OBKMS had been discussed by Egloff et al., 2014.

Furthermore, several tools and systems that deal with the integration of biodiversity and biodiversity data had been developed by different groups. Some of the most important ones are UBio, Global Names, BioGuid, BioNames, Pensoft Taxon Profile, and the Plazi Treatment Repository[5].

Key findings

The key findings from the papers cited in the previous paragraphs can be summarized as follows:

1. Biodiversity science deals with disparate types of data: taxonomic, biogeographic, phylogenetic, visual, descriptive, and others. These data are siloed in unlinked data repositories.

2. Biodiversity databases need a universal system of naming concepts due to the inefficiencies of Linnaean names for modern taxonomy. Taxonomic concept labels have been proposed as a human-readable solution and stable globally unique identifiers of taxonomic concepts had been proposed as a machine-readable solution.

3. There is a base of digitized semi-structured biodiversity information online with appropriate licenses waiting to be integrated as a knowledge base.

Goal and objectives

Given the huge international interest in an open biodiversity knowledge management system (OBKMS), this dissertation started the OpenBiodiv project, the goal of which is to contribute to OBKMS by focusing on biodiversity information extracted from scholarly literature. The scientific goal of the project is to create a formal semantic model of the domain of biodiversity publishing and to apply this model for the creation of a Linked Open Dataset from biodiversity data.

Objective 1: Ontology. Study the domain of biodiversity informatics and biodiversity publishing and develop an ontology allowing data integration from diverse sources.

[5]UBio: http://ubio.org/; Global Names: http://globalnames.org/; BioGuid: http://bioguid.org/; BioNames: http://bionames.org/; Pensoft Taxon Profile: http://ptp.pensoft.eu/; Plazi Treatment Repository: http://plazi.org/wiki/.

Objective 2: Software architecture. Formally define OpenBiodiv as a knowledge-based system and design its integrated software architecture.

Objective 3: Linked open dataset. Create a Linked open dataset (LOD) on the basis of published taxonomic articles using the ontology defined in Objective 1.

Objective 4: Library. Develop methods for converting taxonomic publications into the semantic model of the ontology in order to support Objective 3.

Objective 5: Workflows. Develop practical workflows for continuously converting taxonomic data into taxonomic publications and thus updating the LOD dataset.

Objective 6: Web portal. Create a web portal and example applications on top of the knowledge base.

Methodology

In this section the "meta-choices" (methodological choices) are outlined that have been made before the design and implementation phase. They include but are not limited to programming methodologies and database paradigms.

Choice of database paradigm for OpenBiodiv

We specify OpenBiodiv as a knowledge-based system with a focus on structuring and inter-linking biodiversity data. Two of the possible database technologies that fit this requirement are semantic graph databases (triple stores) such as GraphDB (Ontotext, 2018) and labeled property graphs such as Neo4J (Neo4J Developers, 2012).

Semantic graph databases offer a very simple data model: every fact stored in such a database is composed as a triple of *subject*, *predicate*, and *object*. Subjects of triples are always resource identifiers, whereas objects can be other resource identifiers or literal values (e.g. strings, numbers, etc.). Relations between resources or between resources and literals are given by the predicates (also specified as identifiers). Such links are sometimes referred to as *properties*. Thus, one can visualize a directed graph whose nodes are the subjects or objects, as specified via resource identifiers or literals, and whose arcs are predicates.

Semantic graph databases have the unique feature that the logic layer is also expressed as triples stored in the database. This logic layer, known as *ontology*, is not only responsible for drawing conclusions from the data (inference), but also specifies the semantics of how knowledge should be expressed.

Labeled property graphs, on the other hand, offer a freer data model by allowing the arcs of the knowledge graph to have properties as well. For example, in a labeled property graph whose nodes are two cities A and B and are connected by a property-predicate *road to*, it is possible to additionally attach the value “500 km” to that property. Thus, we indicate that the length of the road connecting the cities is 500 km (Fig. 1).

Note that labeled property graphs are not any more expressive than what can be achieved by triples alone. In fact, complex relationships in a simple triple store can be expressed by making relationships into nodes that have properties on their own. This process is known as *reification*. For example, the two cities A and B can connect to a further node, R indicating the road. R will then have three properties: *start*, *end*,

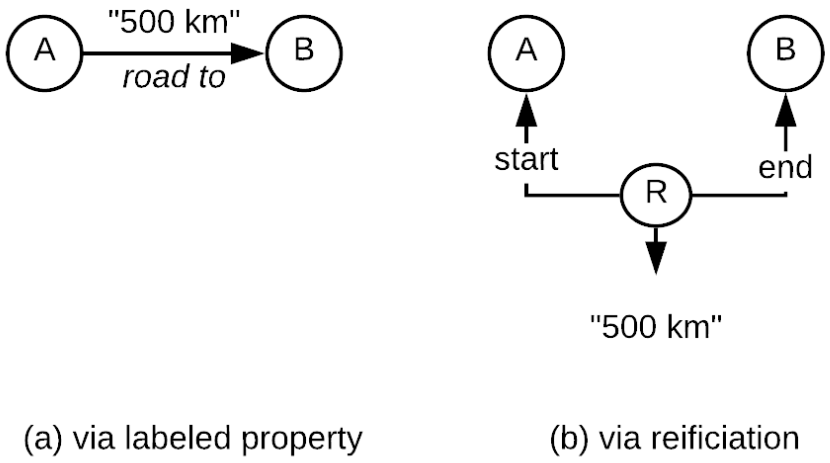


FIGURE 1: Representing the statement that two cities, A and B, are 500 km apart via (a) a labeled property graph and via (b) reification in a semantic graph.

and *length*. The value (object) of *start* will be A, of *end* will be B, and of length will be the literal "500 km" (Fig. 1).

We have summarized the differences between labeled property graphs and semantic graph databases in Table 1. After careful considerations, we settled on the triple store, i.e. semantic graph database as a choice of database technology. This decision was informed by the wide availability of high-quality ontologies and RDF data models in our domain (Baskauf and Webb, 2016; Peroni, 2014) and the popularity of the Semantic Web (Berners-Lee et al., 2001) in the community.

However, we believe that labeled property graphs are a freer and a more natural data model and are perfectly suited for biodiversity informatics. In particular they provide a much more natural formalism for relationships between taxonomic concepts (discussed in Chapter 2). Also, non-RDF semantic databases such as WikiData are gaining in popularity. Therefore, we believe that the applicability of RDF triple stores for OpenBiodiv should constantly be reevaluated.

Choice of information sources

According to *pro-iBiosphere project final report* 2014, biodiversity and biodiversity-related data have two different "life-cycles." In the past, after an observation of a living organism had been made, it was recorded on paper and then the observation record was published in paper-based form. In order for biodiversity data to be available to the modern scientist, efforts are made nowadays to digitize those paper-based publications by Plazi Agosti et al., 2007 and the Biodiversity Heritage Library (Miller et al., 2012). For this purpose, several dedicated XML schemata have been developed (see Penev et al., 2011 for a review), of which TaxPub (Catapano, 2010) and TaxonX seem to be the most widely used (Penev et al., 2012). The digitization of publications contains

TABLE 1: Differences between semantic graph databases (e.g. GraphDB) and labeled property graphs (e.g. Neo4j).

Criterion	Semantic database	Labeled property graph
Semantics	Stored in the database itself as OWL or RDFS statements. Provides a uniform data space. Requires expert ontologists to extract knowledge.	Formal semantics usually are missing. Quick deployment. Uniform data space harder to achieve.
Inference	Provided by the database itself from its ontology or expressed as SPARQL queries. General purpose, slower.	External to the database. Needs to be written for every specific task. Special purpose. Faster.
Community	Has a rich and mature community of ontologists and knowledge engineers. Lots of domain ontologies. Designed for inter-operability. Standards-driven.	Data models are created ad-hoc by data scientists or programmers for a particular task. Inter-operability requires effort and not of primary concern. Applications-driven.

several steps. After scanning and optical character recognition (OCR), text mining is combined with searching for particular kinds of data[6].

In present day, biodiversity data and publications are mostly "born digital" as semantically Enhanced Publications (EP's, Claerbout and Karrenbach, 1992; Godtsenhoven et al., 2009; Shotton, 2009). According to Claerbout and Karrenbach, 1992, *"an EP is a publication that is enhanced with research data, extra materials, post publication data and database records. It has an object-based structure with explicit links between the objects. An object can be (part of) an article, a data set, an image, a movie, a comment, a module or a link to information in a database."* Semantically enhanced publications are thus natives of the Web and the Semantic Web unlike their paper-based equivalents.

The act of publishing in a digital, enhanced format, differs from the ground up from a paper-based publication. The main difference is that a digitally-published document can be structured in such a format as to be suitable both for machine processing and to the human eye. In the field of biodiversity science, Pensoft journals such as ZooKeys, PhytoKeys, and the Biodiversity Data Journal (BDJ) already function by providing EP's (Penev et al., 2010b).

Given the fact that Pensoft Publishers' and Plazi's publications cover a large part of taxonomic literature both in volume and also in temporal span, and the fact that the publications of those two publishers are available as semantic EP's, we've chosen Pensoft's journals and Plazi's treatments as our main sources of information.

Furthermore, we incorporate the taxonomic backbone of GBIF GBIF Secretariat, 2017 as a source for data integration. This is further discussed in Chapter 3.

[6]This procedure leaves a trace in the form of marked-up (tagged) elements that can then be extracted and made available for future use and reuse (Miller et al., 2015).

Choice of development methodology and programming environment

In 2016, based on the outcomes of pro-iBiosphere and on the previous work in the area of biodiversity informatics, we published the Ph.D. plan for this research (Senderov and Penev, 2016). This publication can be considered as the first design specification of OpenBiodiv. However, in the course of developing the system, its design was changed iteratively through a feedback loop from collaborators from the BIG4 project[7] and various international collaborators. We view this positively and in the spirit of both *open science* and *agile software development* (Beck et al., 2001). This iterative approach differs from the waterfall approach where after a through design phase, the specifications "are frozen" and a lengthy implementation phase.

In recent years, the R programming language has been used widely in the field of data science (R Core Team, 2016). R has a rich library of software packages including such for processing XML (Wickham et al., 2018b), for accessing rest API's (Wickham, 2017), and focuses on open science (Boettiger et al., 2015). The capabilities of R as function-oriented and interpreted language allow the iterative software development approach outlined in the previous paragraph to proceed rapidly. Furthermore, R is widely adopted in the biodiversity informatics community. For this reason, the R software environment was chosen as the main programming environment.

Open Science and The Semantic Web

After having specified the desired design and given the programming language, R, we would like to discuss some methodologies and frameworks that have been adopted to be more efficient, open, and reproducible.

We believe that OpenBiodiv needs to be addressed from the point of view of *Open Science*. According to Kraker et al., 2011 and to *Was ist Open Science?*, the six principles of open science are: open methodology, open source, open data, open access, open peer review, and open educational resources. It is our belief that the aim of open science is to ensure access to the whole research product: data, discoveries, hypotheses, and so on. This opening-up will ensure that the scientific product is reproducible and verifiable by other scientists (Mietchen, 2014). There is a very high interest in development of processes and instruments enabling reproducibility and verifiability, as can be evidenced for example by a special issue in Nature dedicated to reproducible research (*Challenges in irreproducible research* 2010). Therefore, the source code, data, and publications of OpenBiodiv will be published openly.

Moreover, OpenBiodiv should be thought of as integral part of the Semantic Web (Berners-Lee et al., 2001). The Semantic Web is a vision for the future of the web where not only documents but also data are connected.

Structure of the thesis

So far the raison d'être of the system and this thesis and an outline of its goal and objectives have been given in this Introduction.

In Chapter 1, a formal specification and design of the desired system as well as an outline of its architecture will be presented; this chapter forms Objective 2 but it is logically convenient to begin the dissertation with it. The subsequent chapters discuss the implementation of OpenBiodiv. Chapter 2 gives a conceptualization of the domain

[7]The Ph.D. candidate, Viktor Senderov, is part of the Marie Skłodowska-Curie BIG4 International Training Network: Biosystematics, informatics and genomics of the big 4 insect groups: training tomorrow's researchers and entrepreneurs.

of scientific taxonomic publishing and formalizes it by introducing the central result of this thesis, the ontology of OpenBiodiv (OpenBiodiv-O) and thus forms Objective 1. Chapter 3 describes the Linked Open Dataset that has been generated based on OpenBiodiv-O and forms Objective 3. Chapter 4 describes in detail the RDF4R software package (an R package for working with RDF), which was used to create the Linked Open Data (OpenBiodiv-LOD) and forms Objective 4. In Chapter 5, two case-studies for importing data into OpenBiodiv from important international repositories are discussed and thus it forms Objective 5. Chapter 6 discusses the website that has being prepared to serve on top of OpenBiodiv-LOD and its applications (Objective 6). In the Conclusion, the results of the dissertation are summarized, the scientific and applied contributions are highlighted, and a discussion of the publications and the dissemination of the results is carried out.

Chapter 1

Architecture of OpenBiodiv

As stated in the Introduction, the goal of the present Ph.D. effort is *"to create an open knowledge-based system of biodiversity information extracted from scholarly literature."* Biodiversity data is quite heterogeneous and comes from many sources; for example, there is taxonomic data (data about the names and descriptions of species), biogeographic data (data about occurrences of organisms at specific locations), genomic data (data about the genetic makeup of species) and so on. For a detailed domain conceptualization including a full discussion of the types of biodiversity data, please refer later to Chapter 2. Due to this heterogeneity and in order to ensure the feasibility of the project as a Ph.D. thesis, the OpenBiodiv system that was developed is focused primarily on creating the models and infrastructure needed for processing scholarly publications of biological systematics and taxonomy.

As per the publication *Final OBKMS Brochure* 2014, the system ought to meet criteria such as providing *"a consistent biodiversity information space," "new formats to support novel and diverse uses," "linkages with other resources," "accreditation for researcher's work,"* and others. Deliberations about the system were published in in the pro-iBiosphere final report (*pro-iBiosphere project final report* 2014). However, the language of the report is high-level and does not provide a formal specification for the system but rather a set of recommendations on the features and implementation of the system. For this reason, at the onset of the project in attempt for formalize the problem we published the specification and design of OpenBiodiv as Ph.D. project plan (Senderov and Penev, 2016).

During the iterative process of agile software development, we refined and extended the design and specification informed by the implementation process. This chapter should serve, therefore, as an updated version of the Ph.D. project plan and as the current specification and design blueprint for the OpenBiodiv system; subsequent chapters contain discussions of the implementation of particular components of the system.

1.1 What is OpenBiodiv?

The understanding of OpenBiodiv as a knowledge-based system can thus be summarized as follows: OpenBiodiv is a database of interconnected biodiversity information together with a logic and application layers allowing users to not only query the data but also discover additional facts of relevance implied by the data. The primary sources of information in OpenBiodiv are the journals of the academic publisher Pensoft, the taxonomic treatments[1] of Plazi, and the taxonomic backbone of GBIF. In Chapter 3 we explore in detail the data sources and their data models.

[1]For a discussion of what a taxonomic treatment is, please refer to the subsequent Chapter 2.

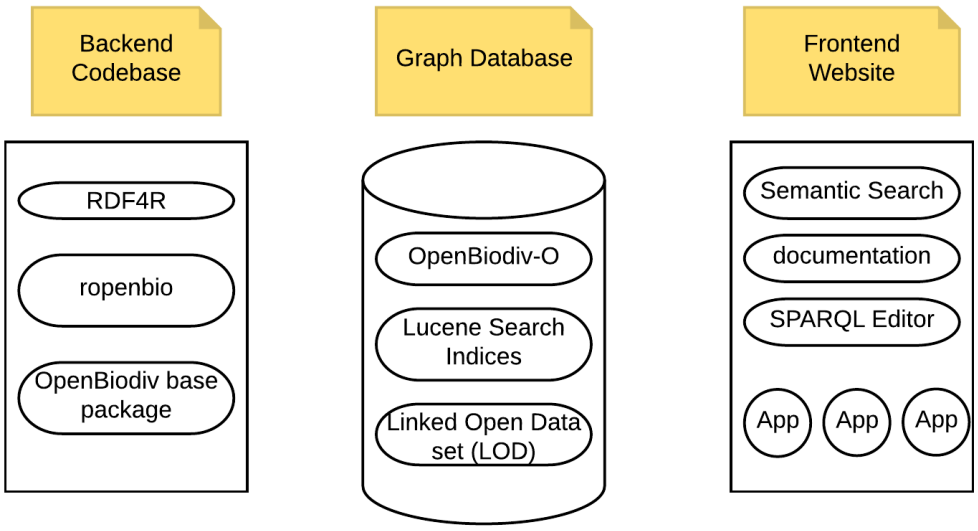


FIGURE 1.1: The components of OpenBiodiv.

The research problem of OpenBiodiv's architecture can be postulated as designing an open access RDF semantic graph database, incorporating information stored in Pensoft, Plazi, and GBIF, and allowing the user of the system to ask complicated queries.

As a blueprint for the type queries in the domain of biodiversity science that should be answerable with the help of the system, we have looked at the list compiled in pro-iBiosphere, 2013. Examples include: *"Is X a valid taxonomic name?" "What are related names to a given name?" "Which authors have published about a given taxon?"*

In this chapter we break up OpenBiodiv into components that will be treated in detail in subsequent chapters. We describe how these components inter-operate in order to former the OpenBiodiv knowledge-based system.

OpenBiodiv consists of (1) a semantic graph database, (2) a code base, and (3) a front-end in the form of a web-portal facilitating the access to the underlying knowledge base (Fig. 1.1). OpenBiodiv enables the flow of information between international repositories for biodiversity data to Biodiversity Data Journal (BDJ) and other journals that use the ARPHA-BioDiv toolkit (Penev et al., 2017a). As a second step, knowledge is extracted from such journals taking advantage of the TaxPub Document Type Definition (DTD)[2] introduced by Catapano, 2010. Example journals include ZooKeys, Biodiversity Data Journal (BDJ), PhytoKeys, MycoKeys, and so on[3]. At the same time, knowledge is extracted from Plazi TreatmentBank, an archive of legacy biodiversity literature published containing over 200 thousand treatments[4] and updated every day. Last but not least, these sources are interlinked via GBIF's taxonomic backbone (GBIF Secretariat, 2017). The extracted knowledge is then stored in a semantic graph database (Fig. 1.2).

[2]We will take the liberty and refer to TaxPub as an XML schema in the rest of the chapter.
[3]The journals can be accessed under `https://pensoft.net/browse_journals`.
[4]A treatment is a special section in a biological publication describing and discussion a species or a higher taxon. TreatmentBank is accessible under `https://pensoft.net/browse_journals`.

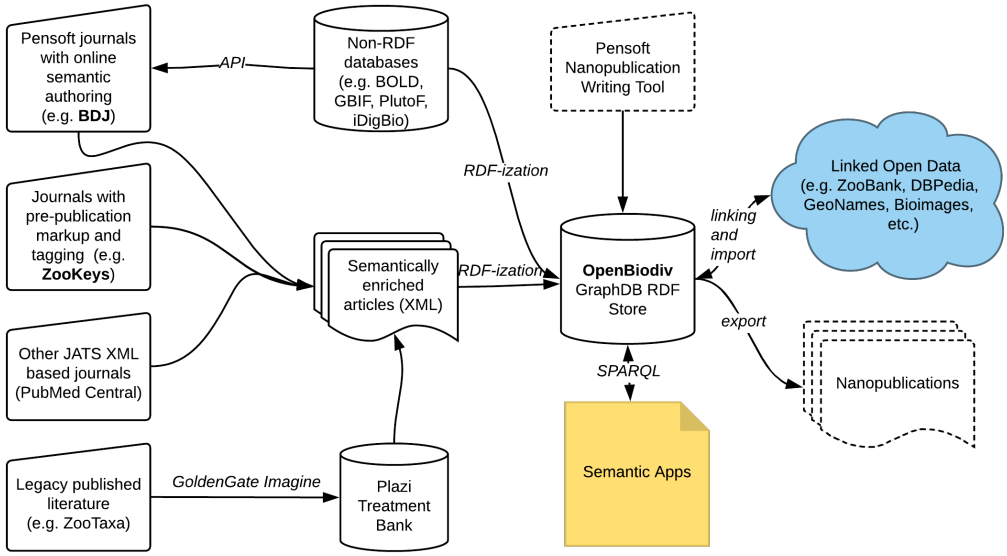


FIGURE 1.2: Flow of information in the biodiversity data space until it reaches the OpenBiodiv semantic database. Dashed lines are components that have not been implemented yet.

1.2 Semantic Graph Database

A primary output of the OpenBiodiv effort is the creation of a semantic database based on knowledge extracted from the archives of Pensoft and Plazi and GBIF's taxonomic backbone and accessible under http://graph.openbiodiv.net/. A discussion of the components of the database follows.

1.2.1 OpenBiodiv ontology (OpenBiodiv-O)

The central result of the OpenBiodiv effort is the creation of a formal domain model of biodiversity publishing, the ontology OpenBiodiv-O (Senderov et al., 2017). The source code of the ontology and accompanying documentation can be accessed under https://github.com/pensoft/openbiodiv-o. A detailed discussion is presented in Chapter 2.

1.2.2 OpenBiodiv Linked Open Dataset (OpenBiodiv-LOD)

Using OpenBiodiv-O and the infrastructure described later in this chapter a dataset incorporating approximately 200 thousand Plazi treatments, 5000 Pensoft articles, as well as GBIF's taxonomic backbone (over a million names) has been created. The dataset is available online through the workbench of the semantic database http://graph.openbiodiv.net. It is discussed in detail in Chapter 3.

1.3 Backend

In order to populate a semantic database it is necessary to create the infrastructure that converts raw data (text, images, data tables, etc.) into a structured semantic format allowing the interlinking of resource identifiers and the answering of complex

queries. OpenBiodiv creates new infrastructure and extends existing infrastructure for transforming biodiversity scholarly publications into Resource Description Format (RDF) statements with the help of the components described in this section.

1.3.1 RDF4R: R package for working with RDF

One of the greater technical challenges for OpenBiodiv is the transformation of biodiversity information (e.g. taxonomic names, paper metadata, figures, etc.) stored as semi-structured XML into fully-structured semantic knowledge in the form of RDF. In order to solve this challenge, an R package has been developed that enables the creation, manipulation, and submission and retrieval to and from a semantic database of RDF statements. This package is accessible under an open source license on GitHub under `https://github.com/pensoft/rdf4r`. We describe the package in Chapter 4.

1.3.2 OpenBiodiv documentation and `ropenbio`

In combination with the RDF4R package, the code-base is completed by one more R package, `ropenbio` and the OpenBiodiv base package of scripts and documentation necessary to bootstrap the database. The package `ropenbio` utilizes the RDF4R package to convert semi-structured XML to RDF. It contains the "mappings" necessary for that conversion. It is available under `https://github.com/pensoft/ropenbio`. The OpenBiodiv base package coordinates the invocation of `ropenbio`, contains scripts for the automatic import of new resources, other housekeeping details, and extensive documentation. It is available under `https://github.com/pensoft/openbiodiv`. The usage of these packages to generate the OpenBiodiv-LOD is discussed in Chapter 3.

1.3.3 Workflow for converting ecological metadata to a manuscript

Ecological Metadata Language (EML) is a popular format for describing ecological datasets (Michener et al., 1997). Biodiversity repositories such as GBIF and DataOne make use of this format to describe the datasets that they store. An import pipeline for importing an EML file as a BDJ data paper[5] has been developed as part of OpenBiodiv (Senderov et al., 2016). We describe this workflow in detail in Chapter 5.[6]

1.3.4 Workflow for importing specimen data into Biodiversity Data Journal

One of the important types of biodiversity data is occurrence data—data that documents the presence of a properly taxonomically identified organism at a given location and time. Such data is stored at international repositories such as BOLD, GBIF, PlutoF, and iDigBio. In order to facilitate data publishing, as well as to act as an entry point into OpenBiodiv, a pipeline for importing any occurrence record from these

[5] A data paper (Chavan and Penev, 2011) is a paper in a scholarly (peer-reviewed) journal discussing a scientific dataset.

[6] To access the pipeline interactively, go to `https://arpha.pensoft.net`, login to the system (registration is free), select “Start a new manuscript,” scroll all the way down to “Import a manuscript,” and follow the necessary steps to upload an EML and use it as a template for your new manuscript.

databases into a BDJ taxonomic paper has been developed (Senderov et al., 2016). We describe this workflow in detail in Chapter 5.[7]

1.4 Front-end

In addition to providing a searchable database endpoint, a website allowing semantic search and containing specific tasks packaged as apps is being developed (`http://openbiodiv.net`). The development of the site extends beyond the scope of the dissertation thesis and is driven by the Pensoft development team. A beta version is already operational Fig. 1.3. A limited discussion is found in Chapter 6.

1.5 IT

The system is deployed on a Debian GNU+Linux virtual machine. GraphDB runs with a 20 GB heap file and with the RDFS-Plus Optimized rule set .This is necessitated by the fact that a performance bottleneck was reached using full OWL inference. These performance issues are discussed in Chapter 3. Continuous operation is ensured by the automatic execution of scripts from the `run` directory of OpenBiodiv Base.

1.6 Discussion and Conclusion

The design of the system began in the second half of 2015 when different database alternatives (Neo4J, GraphDB, WikiBase) and various technologies for storing and serializing RDF were discussed. In addition, a selection was made of sources of information and types of data. The main available data models and ontologies were examined. Following this analysis, we published the system specification and design in Senderov and Penev, 2016 as a Ph.D. project plan. However, during the implementation, the initial plan did not meet the changing requirements of the system and the new challenges that emerged during the implementation. For this reason, in the second and third year of our efforts to create OpenBiodiv we moved from the waterfall development model—where an initial stage of specifying and designing the software architecture is followed by a long-term implementation and testing phase—to the Agile software development (Beck et al., 2001) model, where the specification is broken-down into small "user-stories" that are delivered ad-hoc within one- or two-month sprints. For example, for the RDF4R software library development, some of the user-stories are "Create a framework to work with resource identifiers and literals", "be able to import RDF through the GraphDB API endpoint", etc. The initial fear that this would lead to poorly organized software architecture proved to be unwarranted, because in solving the problems one after another, we were able to isolate them from each other and concentrate on the issues in a consistent manner. Aspects of the Agile software development methodology that we were unable to take full advantage of were the collaborative aspects. For this reason, we did not practice most rituals,

[7]To access the workflow interactively, go to `https://arpha.pensoft.net`, login to the system (registration is free), select "Start a new manuscript," select "Biodiversity Data Journal" as a journal and "Taxonomic Paper" as paper-type and "Create a manuscript." Then, in your new manuscript, expand the "Taxon treatments" section by clicking on the + sign next to it, give a test classification to your treatment (e.g. "Animalia"), click "Save" and you will be presented with a choice of subsections. Click the "Materials" section on the left to visualize the workflow. Look at the lower-part of the dialog, where you see "You may place multiple ID's...". This is the part where you select external resource identifiers to be imported to your article.

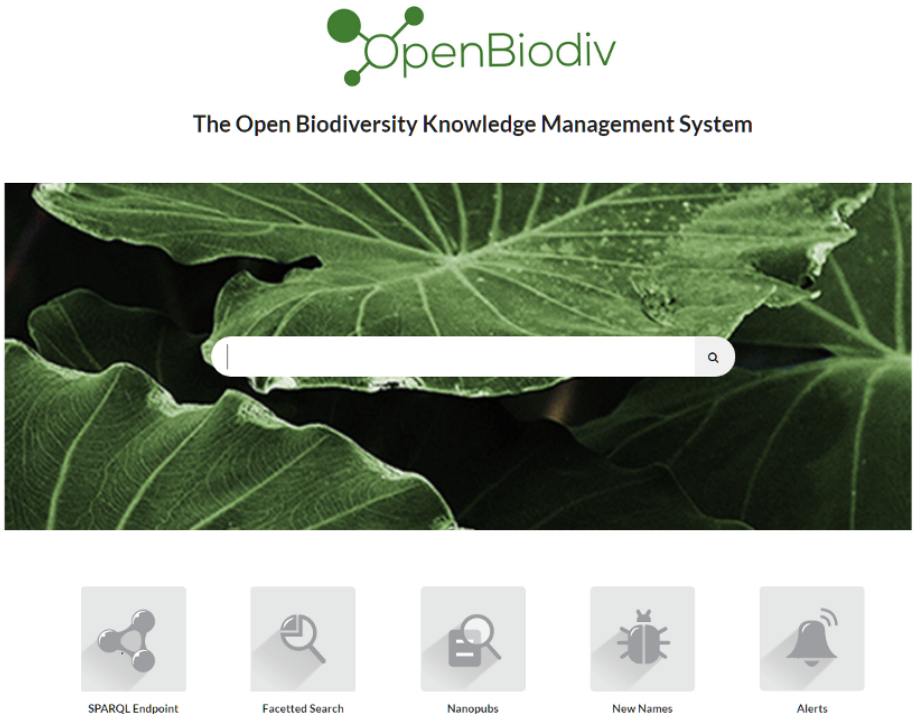


FIGURE 1.3: Beta version of the OpenBiodiv website together with sample app icons.

such as stand-ups and retrospections, and we mainly focused on the iterative software development. Our vision for the future of the system is establish an Agile team to maintain it that takes advantage of the full arsenal of methodology.

Chapter 2

The OpenBiodiv Ontology

OpenBiodiv lifts biodiversity information from scholarly publications and academic databases into a computable semantic form. In this chapter, we introduce OpenBiodiv-O (Senderov et al., 2018), the ontology forming the knowledge and inferencing model of OpenBiodiv. OpenBiodiv-O provides a conceptual model of the structure of a biodiversity publication and the development of related taxonomic concepts. We first introduce the modeled domain in Domain Conceptualization and then formalize it in Results.

By developing an ontology focusing on biological taxonomy, our intent is to provide an ontology that fills in the gaps between ontologies for biodiversity resources such as Darwin-SW and semantic publishing ontologies such as the ontologies comprising the SPAR Ontologies. We take the view that it is advantageous to model the taxonomic process itself rather than any particular state of knowledge.

The source code and documentation are available under the CC BY license[1] from GitHub[2]. We start by introducing the domain of biological taxonomy and the related biodiversity sciences.

2.1 Domain Conceptualization

Biological taxonomy is a very old discipline dating back possibly to Aristotle, whose fundamental insight was to group living things in a hierarchy (Manktelow, 2010). The discipline took its modern form after Carl Linnaeus (1707-1778). In his *Systema Naturae* Linnaeus proposed to group organisms into *kingdoms, classes, orders, genera*, and *species* bearing latinized scientific names with a strictly prescribed syntax. Linnaeus listed possible alternative names and gave a characteristic description of the groups (Linnaeus, 1758). These groups are called *taxa*, which is a Greek word for *arrangement*. The hierarchy that taxa form is called *taxonomy*. The etymology of the word is Greek and roughly translates to *method of arranging*. Note the polysemy here: the science of biological taxonomy is called taxonomy as is the arrangement of taxa itself. We believe, however, that it is sufficiently clear from context what is meant by "taxonomy" in any particular usage throughout this thesis.

Even though Linnaeus and his colleagues may have hoped to describe life on Earth during their lifetimes, we now know that there are millions of species still undiscovered and undescribed (Trontelj and Fiser, 2009). On the other hand, our understanding of species and higher-rank taxonomic concepts changes as evolutionary biology advances (Mallet, 2001). Therefore, an accurate and evolutionarily reliable description of life on Earth is a perpetual process and cannot be completed with a single project that can be converted into an ontology. Thus, our aim is not to create an ontology capturing a

[1] Creative Commons Attribution 4.0 International Public License.
[2] https://github.com/pensoft/openbiodiv-o

fixed view of biological taxonomy, but to create an ontology of the taxonomic process. The ongoing use of this ontology will enable the formal description of taxonomic biodiversity knowledge at any given point in time. In the following paragraphs, we introduce what the taxonomic process entails and reflect on the resources that need modeling.

An examination of the taxonomic process reveals that taxonomy works by employing the scientific method: researchers examine specimens and, based on the phenotypic and genetic variation that they observe, form a hypothesis (Deans et al., 2012). This hypothesis may be called a taxonomic concept, a potential taxon, a species hypothesis (Berendsohn, 1995), or an operational taxonomic unit (OTU, Sokal, 1963) in the case of a numerically delimited taxon.

A taxonomic concept describes the allowable phenotypic, genomic, or other variation within a taxon by designating type specimens and describing characters explicitly. It is a valid falsifiable scientific claim as it needs to fulfill certain verifiable evolutionary requirements. For example, a species-rank taxonomic hypothesis needs to fit our current understanding of species (species concept, Mallet, 2001). More generally, the aspiration is that species concepts are adequate and give certain tangible criteria for species delimitation. However, valid scientific discussions continue about concept adequacy. The discussions are nuanced because they often draw on different conceptions of the relative weight of certain evolutionary phenomena. This leads to having quite a few different species concepts—morphological, ecological, phylogenetic, genomic, biological, etc. (Mallet, 2001). Nevertheless, if we fix a species concept—let us say we take the biological species concept—we can falsify any given species-rank taxonomic hypothesis against our fixed species concept.

Similarly, hypotheses of higher rank (representing upper levels of the taxonomic hierarchy) also need to fulfill certain evolutionary requirements. For example, a modern genus concept requires all species assigned to it to be descendants of a separate lineage and to form a monophyletic clade.

The ranks (taxonomy hierarchy levels) are not completely fixed. The usage of lower ranks (species, genus, family, order) is governed by international Codes (International Commission on Zoological Nomenclature, 1999; *International code of nomenclature for algae, fungi and plants (Melbourne code)* 2012). In the example of Linnaeus' ranks, each organism is first a member of its species, then genus, then order, then class, and finally kingdom. Which specific ranks a given taxonomic study employs is dependent on the field (e.g. botany vs. zoology), on the particular author, on the level of taxonomic resolution required, as well as on the history of classifying in that particular group.

Once the researchers have formed their concept, it must be published in a scientific outlet (journal or book). The biological Codes put some requirements and recommendations aimed at ensuring the quality of published research but ultimately it is a democratic process guaranteeing that everyone may publish taxonomic concepts provided they follow the rules of the Codes. This means that in order to create a knowledge base of biodiversity, we need to be able to mine taxonomic papers from legacy and modern journals and books.

As a first good approximation, a taxonomic concept is based on a number of specimens or occurrences that are listed in a section usually called "Materials Examined." In general terms, we can say that a sighting of a living thing, i.e. an organism, at a given location and at a given time is referred to as an occurrence, and a voucher for this occurrence (e.g. the sampling of the organism itself) is referred to as a specimen (Baskauf and Webb, 2016). Moreover, a taxonomic article may include other specialized sections such as the Checklist section, where one may list all taxa (in fact: the

taxonomic concepts for those taxa) for organisms observed in a given region.

Typically, the information content of a treatment consists of several units. First, we have the aforementioned nomenclatural information that pertains to the scientific name—its authorship, etymology, related names, etc. Then, we have the taxonomic concept information that can be considered to have two components, as well: the first one is the intensional component of the taxonomic concept made up mostly of *traits* or *characters*. Traits are an explicit definition of the allowable variation (e.g. phenotypic, genomic, or ecological) of the organisms that make up the taxon. For example, we can define the order of spiders, Araneae, to be the class of organisms that have specialized appendages used for sperm transfer called pedipalps (Platnick, 2001). Knowledge of this kind is found in the Diagnosis, Description, Distribution and other subsections of the treatment.

Non-traditionally delimited taxonomic hypotheses are called *operational taxonomic units* (OTU's). In the case of genomic delimitation, sometimes the concepts are published directly as database entries and not as Code-compliant taxonomic articles (Page, 2016a). A genomic delimitation can, for example, be based on a barcode sequence and on a statistical clustering algorithm specifying the allowable sequence variability that an organism can possess in order to be considered part of the barcode sequence-bearing operational taxonomic unit. However, as, in the general case, we don't have a Linnaean name or a morphological description for an operational taxonomic unit, we refer to it as a *dark taxon* (Page, 2016a). The term "dark" is, however, usually reserved for concepts at lower ranks. Operational taxonomic units are published, for example, in the form of *barcode identification numbers* (BIN's) in the Barcode of Life Data Systems (BOLD, Ratnasingham and Hebert, 2013), or as *species hypotheses* in Unified system for the DNA based fungal species linked to the classification (UNITE, Kõljalg et al., 2013).

The second part of the information content of a taxonomic concept is the ostensive component: a listing of some (but not necessarily all) of the organisms that belong to the taxonomic concept. This information is found in the Materials Examined subsection of the treatment.

Finally, the relationships between taxonomic concepts—simple hierarchical (*is a*) or more fine-grained Region Connection Calculus 5 (RCC-5, Franz and Peet, 2009; Franz et al., 2016b)—can be both intensionally defined in the nomenclature section or ostensively inferred from the Materials Examined. However, given the customary idiosyncrasies of biological descriptions, providing an initial set of RCC-5 relationships for a machine reasoner to work with often requires expert assessment and cannot be easily lifted from the text.

Thus, in order to model the taxonomic process, our ontology models scholarly taxonomic papers, database entries, agents responsible for their creation, treatments, taxonomic concepts, scientific names, occurrence and specimen information, other entities (e.g. ecological, geographical) part-taking in the taxonomic process, as well as relationships among these.

Previous work

In the biomedical domain there are well-established efforts to extract information and discover knowledge from literature (Momtchev et al., 2009; Williams et al., 2012; Rebholz-Schuhmann et al., 2005). The biodiversity domain, and in particular biological systematics and taxonomy (from here on in this thesis referred to as *taxonomy*), is also moving in the direction of semantization of its research outputs (Kennedy et al., 2005; Penev et al., 2010a; Tzitzikas et al., 2013). The publishing domain has been

modeled through the Semantic Publishing and Referencing Ontologies (SPAR Ontologies, Peroni, 2014). The SPAR Ontologies are a collection of ontologies incorporating—amongst others—FaBiO, the FRBR-aligned Bibliographic Ontology (Peroni and Shotton, 2012), and DoCO, the Document Component Ontology (Constantin et al., 2016). The SPAR Ontologies provide a set of classes and properties for the description of general-purpose journal articles, their components, and related publishing resources. Taxonomic articles and their components, on the other hand, have been modeled through the TaxPub XML Document Type Definition (DTD, also referred to loosely as XML schema) and the Treatment Ontologies (Catapano, 2010; Catapano and Morris, 2016). While TaxPub is the XML-schema of taxonomic publishing for several important taxonomic journals (e.g. ZooKeys, Biodiversity Data Journal), the Treatment Ontologies are still in development and have served as a conceptual template for OpenBiodiv-O.

Taxonomic nomenclature is a discipline with a very long tradition. It transitioned to its modern form with the publication of the Linnaean System (Linnaeus, 1758). Already by the beginning of the last century, there were hundreds of terms in use biological systematics (Witteveen, 2015). At present the naming of organismal groups is governed by by the International Code of Zoological Nomenclature (ICZN, International Commission on Zoological Nomenclature, 1999) and by the International Code of Nomenclature for algae, fungi, and plants (Melbourne Code, *International code of nomenclature for algae, fungi and plants (Melbourne code)* 2012). Due to their complexity (e.g. ICZN has 18 chapters and 3 appendices), it proved challenging to create a top-down ontology of biological nomenclature. Example attempts include the relatively complete NOMEN ontology (Dmitriev and Yoder, 2017) and the somewhat less complete Taxonomic Nomenclatural Status Terms (TNSS, Morris et al.).

There are several projects that are aimed at modeling the broader biodiversity domain conceptually. Darwin Semantic Web (Darwin-SW, Baskauf and Webb, 2016) adapts the previously existing Darwin Core (DwC) terms (Wieczorek et al., 2012) as RDF. These models deal primarily with organismal occurrence data.

Modeling and formalization of the strictly taxonomic domain has been discussed by Berendsohn, 1995 and later, e.g., in Franz and Peet, 2009; Sterner and Franz, 2017. Noteworthy efforts are the XML-based Taxonomic Concept Transfer Schema (Taxonomic Names and Concepts Interest Group, 2006) and a now defunct Taxon Concept ontology (DeVries).

2.2 Methods

OpenBiodiv-O is expressed in Resource Description Framework (RDF). At the onset of the project, a consideration was made to use RDF in favor of a more complex data model such as Neo4J's (Senderov and Penev, 2016). The choice of RDF was made in order to be able to incorporate the multitude of existing domain ontologies into the overall model.

To develop the conceptualization of the taxonomic process and then the ontology we utilized the following process: (1) domain analysis and identification of important resources and their relationships; (2) analysis of existing data models and ontologies and identification of missing classes and properties for the successful formalization of the domain.

The formal structure of the ontology is specified by employing the RDF Schema

(RDFS) and the Web Ontology Language (OWL). It is encoded as a part of a literate programming (Knuth, 1984) document in RMarkdown format titled "OpenBiodiv Ontology and Guide"[3]. The statements have been extracted from the RMarkdown file via *knitr* and are provided here as an appendix. It is also possible to request the ontology via Curl from its endpoint with the indication of `content-type: application/rdf+xml`. The vocabularies can be found as additional appendices, Taxonomic Statuses and RCC-5, and on the GitHub page[4].

A dataset (OpenBiodiv-LOD, will be described in detail in the next Chapter) from Pensoft's journals, Plazi's treatments, and GBIF's taxonomic backbone has been generated with OpenBiodiv-O and can be found at the SPARQL Endpoint [5]. The endpoint is also accessible from the website[6], under "SPARQL Endpoint." Demos are available as "Saved Queries" from the workbench.

2.3 Results

We understand OpenBiodiv-O to be the *shared formal specification of the conceptualization* (Gruber, 1993; Obitko, 2007; Staab and Studer, 2009) that we have introduced in Background. OpenBiodiv-O describes the structure of this conceptualization, not any particular state of it.

There are several domains in which the modeled resources fall. The first one is the scholarly biodiversity publishing domain. The second domain is that of taxonomic nomenclature. The third domain is that of broader taxonomic (biodiversity) resources (e.g. taxonomic concepts and their relationships, species occurrences, traits). To combine such disparate resources together we rely on Simple Knowledge Organization Schema (SKOS) Miles and Bechofer. Unless otherwise noted, the default namespace of the classes and properties for this paper is `<http://openbiodiv.net/>`. The prefixes discussed here are listed at the beginning of the ontology source code.

2.3.1 Semantic Modeling of the Biodiversity Publishing Domain

An article as such may be represented by a set of metadata, while its content consists of article components such as sections, tables, figures and so on (Peroni, 2015).

To accommodate the specific needs of scholarly biodiversity publishing, we introduce a new class for taxonomic articles, Taxonomic Article (`:TaxonomicArticle`), new classes for specific subsections of the taxonomic article such as Taxonomic Treatment, Taxonomic Key, and Taxonomic Checklist, and a new class, Taxonomic Name Usage (`:TaxonomicNameUsage`), for the mentioning of a taxonomic name (see next subsection) in an article. These new classes are summarized in Table 2.1.

The classes from this subsection are based on the TaxPub XML Document Type Definition (DTD, also referred to loosely as XML schema, Catapano, 2010), on the structure of Biodiversity Data Journal's taxonomic paper (Smith et al., 2013), and and on the Treatment Ontologies (Catapano and Morris, 2016).

Furthermore, we introduce two properties: *contains* (`:contains`) and *mentions* (`:mentions`). *contains* is used to link parts of the article together and *mentions* links parts of the article to other concepts.

A graphical representation of the relationships between instances of the publishing-related classes that OpenBiodiv introduces is to be found in the diagram in Fig. 2.1.

[3] http://openbiodiv.net/ontology
[4] https://github.com/vsenderov/openbiodiv-o
[5] http://graph.openbiodiv.net/
[6] http://openbiodiv.net/

TABLE 2.1: New biodiversity publishing classes introduced.

Class QName	Comment
`:Treatment`	section of a taxonomic article
`:NomenclatureSection`	subsection of Treatment
`:NomenclatureHeading`	contains a nomenclatural act
`:NomenclatureCitationList`	list of citations of related concepts
`:MaterialsExamined`	list of examined specimens
`:BiologySection`	subsection of Treatment
`:DescriptionSection`	subsection of Treatment
`:TaxonomicKey`	section with an identification key
`:TaxonomicChecklist`	section with a list of taxa for a region
`:TaxonomicNameUsage`	mention of a taxonomic name

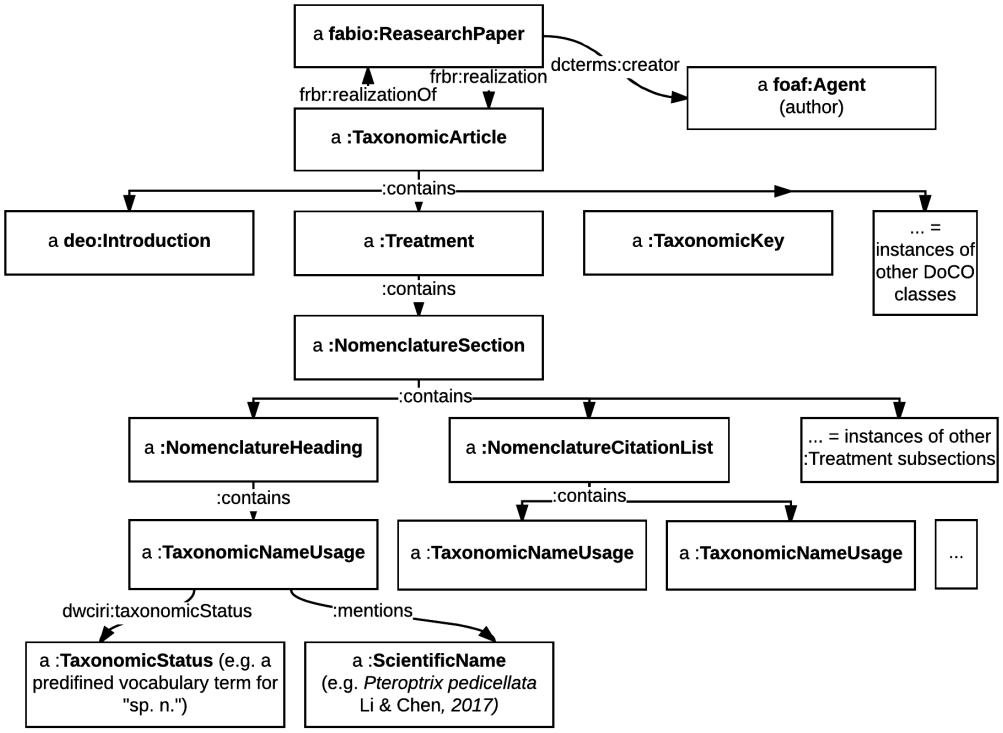


FIGURE 2.1: A graphical representation of the relationships between instances of the publishing-related classes that OpenBiodiv introduces.

Semantics, alignment, and usage

Our bibliographic model has the Semantic Publishing and Referencing Ontologies (SPAR Ontologies) at its core with a few extensions that we have written to accommodate for taxonomic elements. The SPAR Ontologies' FRBR-aligned Bibliographic Ontology (FaBiO) uses the Functional Requirements for Bibliographic Records (FRBR, Tillett, 2003) model to separate publishable items into less or more abstract classes. We deal primarily with the Work class, i.e. the conceptual idea behind a publishable item (e.g. the story of "War and Peace" as thought up by Leo Tolstoy), and the Expression class, i.e. a version of record of a Work (e.g. "War and Peace," paperback

edition by Wordsworth Classics).

Taxonomic Article is a subclass of FaBiO's Journal Article. Furthermore Journal Article is a FRBR Expression. This implies that taxonomic articles are FRBR expressions as well. This has important implications later on when discussing taxonomic concept labels. Also, it means that we separate the abstract properties of an article (in a FaBiO Research Paper instance, which is a Work) from the version of record (in a Taxonomic Article, an Expression).

The taxonomic-specific section and subsection classes are introduced as subclasses of Discourse Element Ontology's (DEO) Discourse Element (`deo:DiscourseElement`, Constantin et al., 2016). So is the class Mention (`:Mention`), meant to represent an area of a document that can be considered a mention of something. This class, and the corresponding property, *mentions*, are inspired by `pext:Mention` and its corresponding property from PROTON (Damova et al., 2010). The redefinition is necessary by the fact in OpenBiodiv-O they possess a slightly different semantics and a different placement in the upper-level hierarchy. we then introduce Taxonomic Name Usage as a subclass of Mention.

This placement of the document component classes that we have introduced introduced in Discourse Element means that they ought to be used exactly in the same way as one would use the other discourse elements from DEO and DoCO (analogous to e.g. `deo:Introduction`). Note: DEO is imported by DoCO. Figs. 2.2 and 2.3 give example usage in Turtle illustrating these ideas. A caveat here is that while the SPAR Ontologies use `po:contains` in their examples, we use *contains*, which is a subproperty of `po:contains` with the additional property of being transitive. We believe this definition is sensible as surely a sub-subcomponent is contained in a component. All other aspects of expressing a taxonomic article in RDF according to OpenBiodiv-O are exactly the same as according to the SPAR Ontologies.

2.3.2 Semantic modeling of biological nomenclature

While NOMEN and TNSS (introduced in subsection "Previous work") take a top-down approach of modeling the nomenclatural Codes, OpenBiodiv-O takes a bottom-up approach of modeling the use of taxonomic names in articles. Where possible we align OpenBiodiv-O classes to NOMEN.

Based on the need to accommodate taxonomic concepts, we have defined the class hierarchy of taxonomic names found in Fig. 2.4. Furthermore, we have introduced the class Taxonomic Name Usage (`:TaxonomicNameUsage`). Taxonomic name usages have been discussed widely in the community (e.g. in Pyle, 2016a); however, the meaning of term remains vague. The abbreviation TNU is used interchangeably for "taxon name usage" and for "taxonomic name usage." In OpenBiodiv-O, a taxonomic name usage is the mentioning of a taxonomic name in the text, optionally followed by a taxonomic status.

For example, "*Heser stoevi* Deltschev 2016, sp. n." is a taxonomic name usage. The cursive text followed by the author and year of the original species description is the latinized scientific name. The abbreviation "sp. n." stands for the Latin *species novum*, indicating the discovery of a new taxon.

We also introduce the class Taxonomic Concept Label (`:TaxonomicConceptLabel`). A taxonomic concept label (TCL) is a Linnaean name plus a reference to a publication, where the discussed taxon is circumscribed. The link is via the keyword "sec." (Latin for (*secundum*, Berendsohn, 1995). An example would be "*Andropogon virginicus* var. *tenuispatheus* sec. Blomquist, 1948". Here, Blomquist, 1948 is a valid bibliographic reference to the publication where the concept is circumscribed.

```
:biodiversity-data-journal rdf:type fabio:Journal ;
        skos:prefLabel "Biodiversity Data Journal"@en ;
        skos:altLabel  "BDJ"@en ;
        fabio:issn     "1314-2836" ;
        fabio:eIssn    "1314-2828" ;
        frbr:part      :b90f6933-ab5e-4ce1-9379-12de9ef4eaa6 .

<http://dx.doi.org/10.3897/BDJ.1.e953> rdf:type fabio:TaxonomicArticle ;
        skos:prefLabel           "10.3897/BDJ.1.e953" ;
        dc:title                 "Casuarinicola australis Taylor, 2010
        (Hemiptera: Triozidae), newly recorded from New Zealand"@en ;
        prism:doi                "10.3897/BDJ.1.e953" ;
        dcelements:publisher     "Pensoft Publishers"@en ;
        fabio:hasPublicationYear "2013"^^xsd:gYear ;
        prism:publicationDate    "2013-9-16"^^xsd:date ;
        dcterms:publisher        :pensoft-publishers ;
        frbr:realizationOf       :thorpe-2013 .

:thorpe-2013 rdf:type  :ResearchPaper ;
        skos:prefLabel       "Thorpe 2013"
        skos:altLabel        "paper10.3897/BDJ.1.e953" ;
        dcterms:creator      :stephen-e-thorpe ;
        prism:keywords       "Casuarinicola australis"@en ;
        fabio:hasSubjectTerm :a2ee4929-90dd-4a7a-aa5c-08836f49d549 .

:pensoft-publishers rdf:type :Publisher ;
        skos:prefLabel "Pensoft Publishers"@en .

:stephen-e-thorpe rdf:type foaf:Person ;
        skos:prefLabel "Stephen E. Thorpe" ;
        foaf:firstName "Stephen E." ;
        foaf:surname   "Thorpe" ;
        foaf:mbox      "stephen_thorpe@yahoo.co.nz" ;
        :affiliation   "School of Biological Sciences (Tamaki Campus),
              University of Auckland, Auckland, New Zealand"@en .

:a2ee4929-90dd-4a7a-aa5c-08836f49d549 rdf:type fabio:SubjectTerm ;
        rdfs:label    "Casuarinicola australis"@en ;
        skos:inScheme :openbiodiv-subject-terms .
```

FIGURE 2.2: This example shows how to express the metadata of a taxonomic article with the SPAR Ontologies' model and the classes that OpenBiodiv defines. The code is in Turtle.

We extracted taxonomic status abbreviations from about 4,000 articles across four taxonomic journals (ZooKeys, Biodiversity Data Journal, PhytoKeys, and MycoKeys) in order to create a taxonomic status vocabulary (see appendices) that covers the eight most common cases (Table 2.2). The Latin abbreviations that have been classified into these classes can be found on the OpenBiodiv-O GitHub page. (See Methods for more details).

TABLE 2.2: OpenBiodiv Taxonomic Status Vocabulary.

Vocabulary Instance QName	Example Abbrev	Comment
`:TaxonomicUncertainty`	*incertae sedis*	Taxonomic Uncertainty
`:TaxonDiscovery`	*sp. n.*	Taxonomic Discovery
`:ReplacementName`	*comb. n.*	Replacement Name
`:UnavailableName`	*nomen dubium*	Unavailable Name
`:AvailableName`	*stat. rev.*	Available Name
`:TypeSpecimenDesignation`	*lectotype designation*	Type Specimen Designation
`:TypeSpeciesDesignation`	*type species*	Type Species Designation
`:NewOccurrenceRecord`	*new country record*	New Occurrence Record (for region)

```
<http://dx.doi.org/10.3897/BDJ.1.e953>
  :contains :abstract, :casuarinicola-australis-treatment .

:introduction rdf:type deo:Introduction, doco:Section ;
  c4o:hasContent "Casuarinicola australis Taylor, 2010 was described from
        Australia, where it is the most common and widespread member of its
        genus, being widely distributed in New South Wales, Queensland,
        South Australia, Victoria and Western Australia. "

:casuarinicola-australis-treatment rdf:type doco:Section, :Treatment ;
  :contains :casuarinicola-australis-nomenclature ,
              :casuarinicola-australis-materials ,
              :casuarinicola-australis-description ,
              :figure-box-1 ,
              :figure-box-2 .

:casuarinicola-australis-nomenclature rdf:type :NomenclatureSection ;
  :contains :casuarinicola-australis-nomenclature-heading .

:casuarinicola-australis-nomenclature-heading a :NomenclatureHeading ;
  cnt:chars "Casuarinicola australis Taylor, 2010" .

:casuarinicola-australis-materials rdf:type :MaterialsExamined ;
  c4o:hasContent "country: New Zealand;
                  verbatimLocality: Mechanics Bay, Auckland City;
                  verbatimElevation: 0-5 m;
                  verbatimLatitude: 36.84749381055 ;
                  verbatimLongitude: 174.7869624545E ;
                  eventDate: 6 January 2013;
                  sex: 1 male, 1 female;
                  recordedBy: Stephen Thorpe;
                  institutionCode: Auckland Museum" .

:casuarinicola-australis-description rdf:type :DescriptionSection ;
  c4o:hasContent "On 6 Jan 2013, I examined some Casuarina glauca trees growing
        in the vicinity of Ports of Auckland at Mechanics Bay." .
```

FIGURE 2.3: This examples shows how to express the article structure with the help of `:contains`. The code is in Turtle.

Based on our analysis of taxonomic statuses, we have identified two Code-compliant patterns of relationship between latinized scientific names (Fig. 2.5). The pattern *replacement name*, implemented via the property `:replacementName`, indicates that a certain Linnaean name should be used instead of another Linnaean name. It covers a wide variety of cases in the Codes, such as, for example, the placement of one species taxon in a new genus ("comb. n."), the correction of a name for nomenclatural reasons ("nomen novum"), or the application of the Principle of Priority for the discovery of synonyms ("syn. nov.", International Commission on Zoological Nomenclature, 2017).

The other pattern is that of *related names* (`:relatedName`). It is a broader pattern, indicating that two names are somehow related. For example, they may be synonyms, with one replacing the other, or they may point to taxonomically related taxonomic concepts. For example, *Harmonia manillana* (Mulsant, 1866) is related to *Caria manillana* Mulsant, 1866 since, as per Poorani and Booth, 2016, a name-bearing type (lectotype) of *Harmonia manillana* (Mulsant, 1866) sec. Poorani Poorani and Booth, 2016 is named *Caria manillana* Mulsant, 1866.

Semantics, alignment and usage

As evident from Fig. 2.4, OpenBiodiv-O taxonomic names are aligned to NOMEN names.

The linking between text and taxonomic names must pass through the intermediary class Taxonomic Name Usage. As parts of the manuscript, taxonomic name

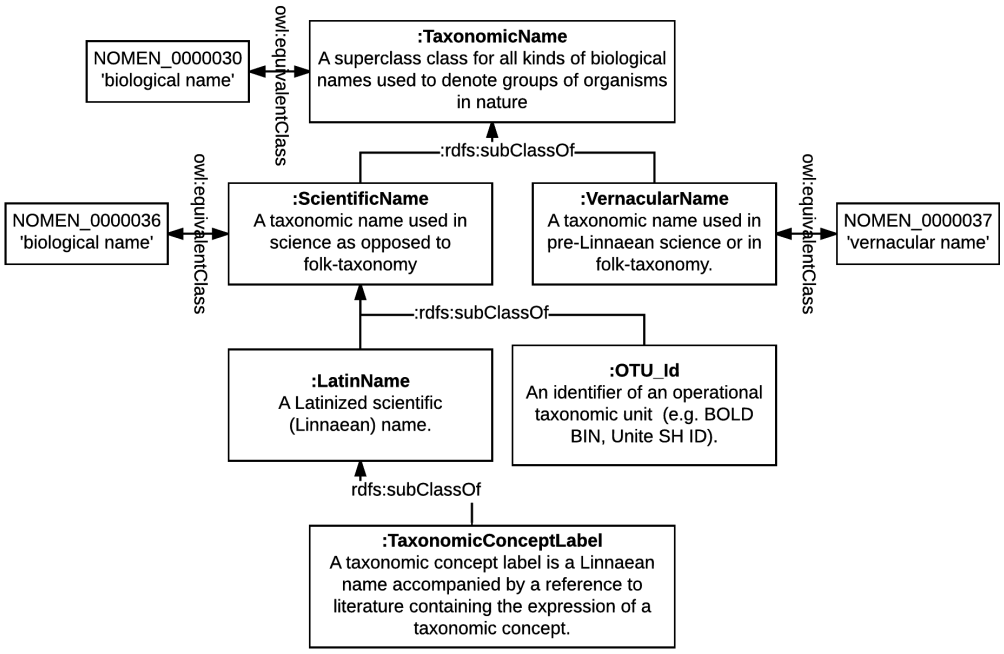


FIGURE 2.4: We created this class hierarchy to accommodate both traditional taxonomic name usages and the usage of taxonomic concept labels and operational taxonomic units.

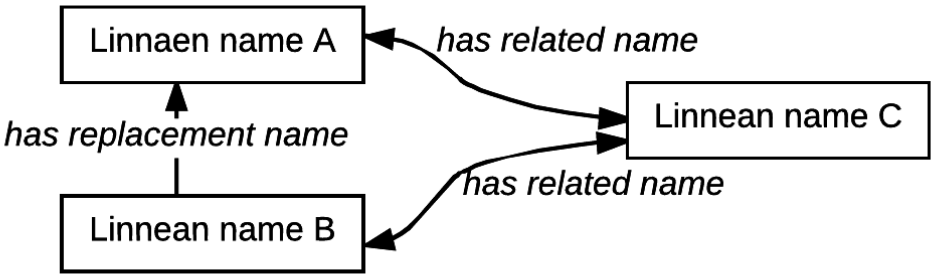


FIGURE 2.5: Chains of *replacement names* can be followed to find the currently used name. *Related name* indicates that two names are related somehow, but not which one is preferable.

usages link document components to taxonomic names. Taxonomic name usages are *contained* in sections such as Treatment, and *mention* a taxonomic name as illustrated in the example in Fig. 2.6.

2.3.3 Semantic Modeling of the Taxonomic Concepts

In OpenBiodiv-O taxonomic names are not the carriers of semantic information about taxa. This task is accomplished by a new class, Taxonomic Concept (`:TaxonomicConcept`).

```
:casuarinicola-australis-nomenclature-heading
  po:contains :casuarinicola-australis-TNU .

:casuarinicola-australis-TNU a :TaxonomicNameUsage ;
  dc:date "2013-9-16"^^:xsd:date ;
  cnt:chars "Casuarinicola australis Taylor, 2010" ;
  dwc:genus "Casuarinicola" ;
  dwc:specificEpithet "australis" ;
  dwc:scientificNameAuthorship "Taylor, 2010" ;
  # we can infer the following because we are in the treatment heading
  dwc:nameAccordingToId "doi: 10.3897/BDJ.1.e953" ;
  pkm:mentions :casuarinicola-australis-taylor,
                          :casuarinicola-australis-taylor-sec-thorpe-2013 .

:casuarinicola-australis-taylor a :ScientificName ;
  rdfs:label "Casuarinicola australis Taylor, 2010" ;
  dwc:genus "Casuarinicola" ;
  dwc:specificEpithet "australis" ;
  dwc:scientificNameAuthorship "Taylor, 2010" .

:casuarinicola-australis-taylor-sec-thorpe-2013 a :TaxonomicConceptLabel ;
  rdfs:label "Casuarinicola australis Taylor, 2010 sec. Thorpe 2013" ;
  dwc:genus "Casuarinicola" ;
  dwc:specificEpithet "australis" ;
  dwc:scientificNameAuthorship "Taylor, 2010" .
  dwc:nameAccordingToId "doi: 10.3897/BDJ.1.e953" ;
  :nameAccordingTo <http://dx.doi.org/10.3897/BDJ.1.e953> .
```

FIGURE 2.6: This examples shows how taxonomic name usages link document components to taxonomic names. The code is in Turtle.

A taxonomic concept is the theory that a taxonomist forms about a taxon in a scholarly biological taxonomic publication and thus always has a taxonomic concept label. We also introduce a more general class, Operational Taxonomic Unit (`:OperationalTaxonomicUnit`) that can be used for all kinds of taxonomic hypotheses, including ones that don't have a proper taxonomic concept label. The class hierarchy has been illustrated in Fig. 2.7.

Taxonomic concepts are related to taxonomic names—including taxonomic concept labels—via the property *has taxonomic name* (`:taxonomicName`) and its sub-properties mimicking in their range the hierarchy of taxonomic names that we introduced earlier. We have defined a property specifically to link taxonomic concepts to taxonomic concept labels, *has taxonomic concept label* (`:taxonomicConceptLabel`). The property hierarchy diagram is shown in Fig. 2.8.

There are two ways to relate taxonomic concepts to each other (Fig. 2.9). As we pointed out earlier, historically taxonomic concepts form the hierarchy known as biological taxonomy. To express such simple semantic relations, it is fully sufficient to use the SKOS semantic vocabulary Miles and Bechofer.

However, these simple relationships are not well suited for machine reasoning. This is why Franz and Peet Franz and Peet, 2009 suggested, building on previous work by e.g. Koperski et al., 2000, to use the RCC-5 language to express relationships between taxonomic concepts. Furthermore, the Euler (Chen et al., 2014) program was developed, which uses Answer Set Programming (ASP) to reason over RCC-5 taxonomic relationships. An answer set reasoner is not part of OpenBiodiv as this task can be accomplished by Euler; however, we have provided an RCC-5 dictionary class (`:RCC5Dictionary`), an RCC-5 relation term class (`:RCC5Relation`), a vocabulary of such terms to express the RCC-5 relationships in RDF (see appendices), as well as a class and properties to express RCC-5 statements (`:RCC5Statement`, `:rcc5Property`,

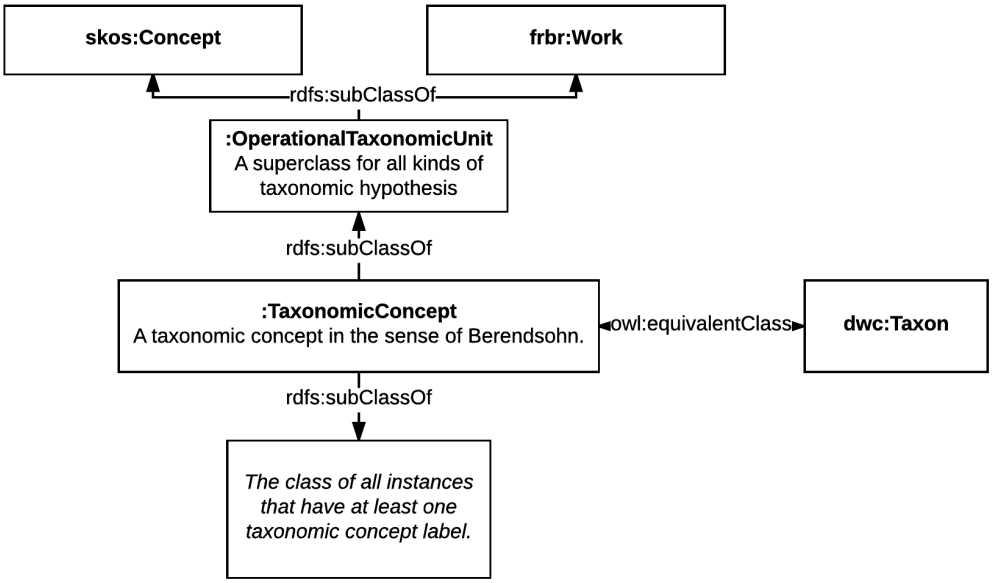


FIGURE 2.7: A taxonomic concept is a `skos:Concept`, a `frbr:Work`, a `dwc:Taxon` and has at least one taxonomic concept label.

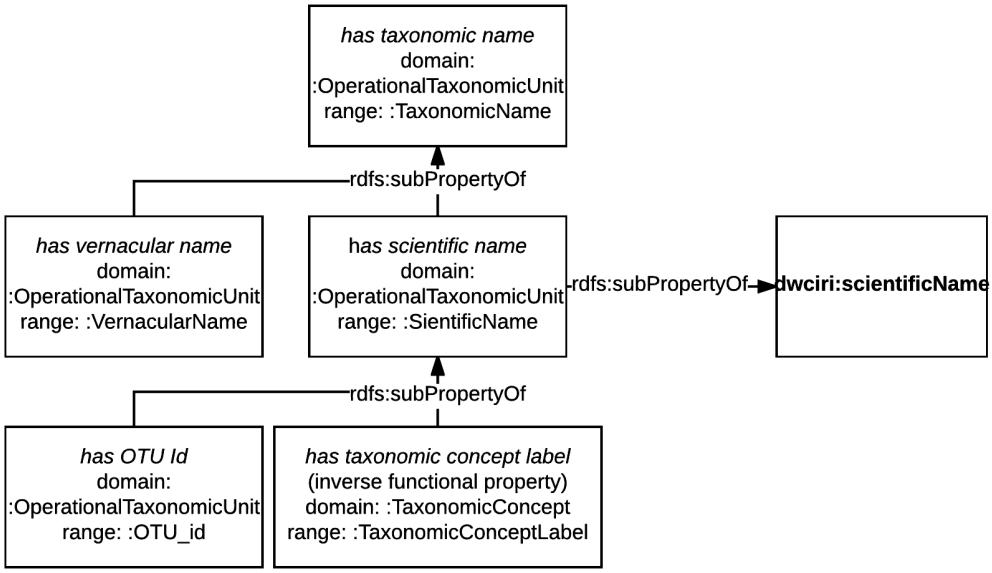


FIGURE 2.8: Property hierarchy is aligned with the taxonomic name class hierarchy and with DarwinCore.

and subproperties).

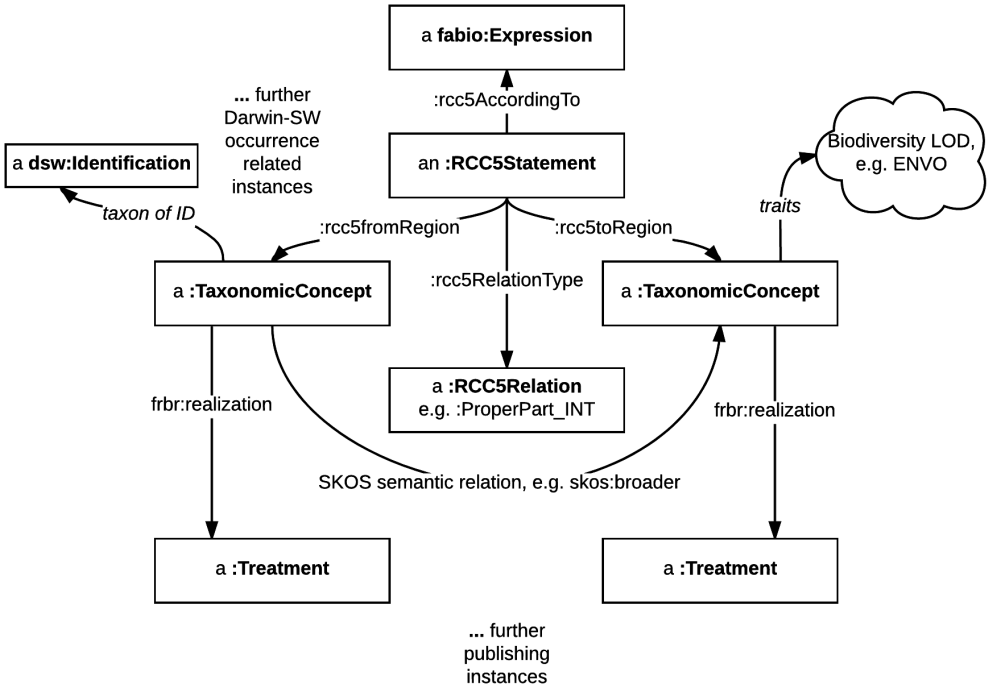


FIGURE 2.9: In order to express an RCC-5 relationship between concepts, create an `:RCC5Sgtatement` and use the corresponding properties to link two taxonomic concepts via it. Further, taxonomic concepts are linked to traits (e.g. ecology in ENVO), occurrences (e.g. Darwin-SW) and realize treatments.

Semantics and alignment

We introduce Taxonomic Concept as equivalent (`owl:equivalentClass`) to the DwC term Taxon (`dwc:Taxon`) [7]. However, by including "concept" in the class' name, we highlight the fact that the semantics it carries reflect the scientific theory of a given author about a taxon in nature. As we mentioned earlier, our ontology models the ongoing still unfinished process of taxonomic discovery. For this reason, we also derive Taxonomic Concept from Work. This derivation fits the definition of Work in FRBR/FaBiO, which is *"a distinct intellectual or artistic creation."* Finally, as we use SKOS to connect taxonomic concepts to each other, we derive Taxonomic Concept from SKOS Concept.

As with other semantic publishing-related aspects of the ontology, the creation of the RCC-5 vocabulary follows the SPAR Ontologies' model. Thus OpenBiodiv RCC-5 Vocabulary (`:RCC5RelationshipTerms`) is a SKOS concept scheme and every RCC-5 Relation is a SKOS concept. This allows to seamlessly share this vocabulary with other publishers of biodiversity information that also follow the SPAR Ontologies' model.

It is important to note that we have aligned the subproperty of *has taxonomic name*, *has scientific name* (`:scientificName`), to the DwC property `dwciri:scientificName`.

[7] A list of the current DwC terms is maintained by TDWG under http://rs.tdwg.org/dwc/terms/

The difference is that while the DwC property is unbound and provides more flexibility, the OpenBiodiv-O property has the domain Taxonomic Concept and the range Scientific Name and provides for inference. Furthermore, *has taxonomic concept label* is an inverse-functional property with the domain Taxonomic Concept. This means that a given taxonomic concept label uniquely determines its taxonomic concept. This is accomplished by a minimum cardinality restriction on the property.

Together with the declaration of *has taxonomic concept label* to be an inverse functional property, we can now list what types of relationships between names and taxonomic concepts are allowed: (1) The relationship between a taxonomic concept and a name that is not a taxonomic concept label is many-to-many—i.e. one Linnaean name can be a mention of multiple taxonomic concepts, and one taxonomic concept may have multiple Linnaean names. (2) The relationship between a taxonomic concept and a taxonomic concept label is one-to-many: while a taxonomic concept may have more than one (at least one is needed) labels, every label uniquely identifies a concept. These logical restrictions make taxonomic concept labels into unique identifiers to taxonomic concepts, something that Linnaean names are not.

Usage

For an example of linking two taxonomic concepts to each other, let us look at the species-rank concept *Casuarinicola australis* Taylor et al., 2010 sec. Thorpe, 2013. It is a narrower concept than the genus-rank concept of *Casuarinicola* Taylor et al., 2010 sec. Taylor et al., 2010. As we have aligned our concepts to SKOS, we can use its vocabulary to express this statement as seen in the example in Fig. 2.10. A further example of how to utilize the OpenBiodiv RCC-5 vocabulary is found in Fig. 2.11.

```
:concept-casuarinicola-australis-thorpe rdf:type :TaxonomicConcept ;
  :taxonomicConceptLabel :casuarinicola-australis-taylor-sec-thorpe-2013 .

:concept-casuarinicola-taylor rdf:type :TaxonomicConcept ;
  skos:broader concept-thorpe .
```

FIGURE 2.10: We can use SKOS semantic properties to illustrate simple relationships between taxonomic concepts.

```
:statement rdf:type :RCC5Statement ;
  :rcc5FromRegion :concept-casuarinicola-australis-thorpe ;
  :rcc5ToRegion :concept-casuarinicola-taylor ;
  :rcc5AccordingTo <http://dx.doi.org/10.3897/BDJ.1.e953> ;
  :rcc5RelationType :ProperPart_INT .
```

FIGURE 2.11: In order to express an RCC-5 relationship between concepts, create an `:RCC5Sgtatement` and use the corresponding properties to link two taxonomic concepts via it. SKOS relations relate concepts directly.

Furthermore, thanks to the alignment to DwC, we treat instances of our class Taxonomic Concept as functionally equivalent to DwC Taxa. This makes linking to other biodiversity ontologies possible. For example, the Open Biomedical Ontologies' (OBO) Population and Community Ontology (PCO, Walls et al., 2014) has a class "collection of organisms" (`http://purl.obolibrary.org/obo/PCO_000000`) that can

be considered a superclass of DwC Taxon. Therefore, every taxonomic concept is a collection of organisms and the application of OBO properties on it is allowed.

In the paper that inspired our *Casuarinicola* example (Thorpe, 2013), we read: *"On 26 February 2013, the species was found to be fairly common on Casuarina trees at Thomas Bloodworth Park, Auckland."* This statement can be interpreted (in ENVO) as meaning that the taxonomic concept that the author formulated implies that it includes the habitat "forest biome" (http://purl.obolibrary.org/obo/RO_0002303). The RDF example is shown in Fig. 2.12.

```
:australian-casuarina-forest rdf:type <http://purl.obolibrary.org/obo/ENVO_01000174> .
:hasHabitat owl:sameAs <http://purl.obolibrary.org/obo/RO_0002303> .
:concept-casuarinicola-australis-thorpe :hasHabitat :australian-casuarina-forest .
```

FIGURE 2.12: We create a shortcut for *has habitat* and instance of the "forest biome" and link them to our taxonomic concept in order to express the fact that specimens of it have been found to live in *Casuarina* trees.

As we pointed out earlier, taxonomic concepts have an intensional component (traits or characters) and an ostensive component (a list of occurrences belonging to the concept). The ostensive component can be expressed by linking occurrences to the taxonomic concepts via Darwin-SW. This is possible as we have aligned the Taxon Concept class to DwC Taxon used by Darwin-SW. For an example refer to Baskauf and Webb, 2016.

Lastly, describing traits is an active area of ontological research (Huang et al., 2015). Due to the very complex language used to describe morphological characteristics, the Ontology Term Organizer (OTO, Huang et al., 2015) software was developed to allow for user-created vocabularies. An avenue for a follow-up project is to work with OTU to express traits and trait equivalences (in the taxonomic sense) during the population of OpenBiodiv with triples (Hong et al., 2018).

Last, the interpretation of Taxonomic Concepts as Work means that they are realized by taxonomic treatments (e.g. Fig 2.13).

```
:casuarinicola-australis-treatment frbr:realizationOf :concept-casuarinicola-australis-thorpe.
```

FIGURE 2.13: A treatment is the realization of a taxonomic concept.

2.4 Discussion

OpenBiodiv-O is—together with the Treatment Ontologies (Catapano and Morris, 2016)—the first effort to model taxonomic articles as RDF. It introduces classes and properties in the domains of biodiversity publishing and biological taxonomy and aligns them with the SPAR Ontologies, the Treatment Ontologies, the Open Biomedical Ontologies (OBO), TaxPub, NOMEN, and DarwinCore. We believe this introduction bridges the ontological gap that we had outlined in our aims and allows for the creation of a Linked Open Dataset (LOD) of biodiversity information (biodiversity knowledge graph, Senderov and Penev, 2016; Page, 2016b).

Furthermore, this biodiversity knowledge graph, together with this ontology, additional semantic rules, and user software will form the OpenBiodiv Knowledge Management System. This system, as any taxonomic information system should, has taxonomic names as a key building block. For any given taxonomic name, the user will be able to rely on two patterns—*replacement name* and *related name*—to get answers to two questions of high importance to the working taxonomist. First: what is the current and historical usage of any given Linnaean name? Second: given a particular name, what other related names ought to be considered in a taxonomic discussion?

Both may be useful in building semantic search applications and the latter, in particular, is actively being researched by a group at the National Center for Text Mining in the UK (NaCTeM, Nguyen et al., 2017). OpenBiodiv-O proper does not include a mechanism for inferring replacement names and related names; however, such mechanisms are part of the OpenBiodiv knowledge system via SPARQL rules using information encoded in the document structure (Nomenclature section). Another way to infer related names is via a machine learning approach to obtain feature vectors of taxonomic names. Note that the ontology can describe related names independent of the process of their generation and will enable the comparison of both approaches in a future work.

On the other hand, by using OpenBiodiv-O, a knowledge-based system does not have to have a backbone name-based taxonomy. A backbone taxonomy is a single, monolithic hierarchy in which any and all conflicts or ambiguities have been pragmatically (socially, algorithmically) resolved, even if there is no clear consensus in the greater taxonomic domain. Such backbone taxonomies are used in systems that rely solely on taxonomic names (and not concepts) as bearers of information. They are needed as it is impossible, in such a system, to express two different sets of statements for a single name.

In OpenBiodiv, however, multiple hierarchies of taxonomic concepts may exist. For example, large synthetic taxonomies such as GBIF's backbone taxonomy (GBIF Secretariat, 2017) or Catalogue of Life may not agree or may have some issues (Page, 2012). With OpenBiodiv-O, we may, in fact, incorporate both these taxonomies at the same time! It is possible according to the ontology to have two sets of taxonomic concepts (even with the same taxonomic names) with a different hierarchical arrangement. By allowing this, we leave some room for human interpretation as an additional architectural layer. Thus, we delay the decision of which hierarchy to use to the user of the system (e.g. a practicing taxonomist) and not to the system's architect. Due to this design feature, it is likely that our system stands a better chance to be trusted as a science process-enabling platform as the system architects don't force a taxonomic opinion on the practicing taxonomist.

It should be noted that a successful concept-based system exists for the taxonomic order Aves (birds) (Lepage et al., 2014). The main issue that we will face is to develop tools to enable expert users to annotate taxonomic concepts with the proper relationships as only recently individual articles utilizing concept taxonomy in addition to nomenclature have been published (Franz et al., 2016b; Jansen and Franz, 2015; Franz and Zhang, 2017). We do believe that their numbers will rise driven by the realization that there are some problems with relying solely on Linnaean names for the identification of taxonomic concepts (Patterson et al., 2010; Remsen, 2016; Franz et al., 2016a). Concept taxonomy may, in fact, become even more important in the future as conservation efforts face challenges due to unresolved taxonomies (Garnett and Christidis, 2017). Properly aligning taxonomic concepts to nomenclature across revisions (Franz et al., 2016c) may be the solution.

Together with taxonomic information, the ontology allows modeling the source information in a knowledge base. This will be useful for meta-studies, for the purposes of reproducible research, and other scholarly purposes. Moreover, it will be an expert system as the knowledge extracted will come from scholarly publications. We envision the system to be able to address a wide variety of taxonomic competency questions raised by researchers (pro-iBiosphere, 2013). Examples include: *"Is X a valid taxonomic name (in a nomenclatorial sense)?" "Which treatments use different names for the same taxon concepts?" "Which treatments are nomenclatorially linked (including homonyms!) to another treatment?"*

In the next Chapter we will show how we populated the ontology with triples extracted from Pensoft journals, legacy journals text-mined by Plazi, as well as databases such as GBIF's taxonomic backbone (GBIF Secretariat, 2017). Special effort will be made to link the dataset to the Linked Open Data cloud via resources such as geographic or institution names. In terms of extending the ontological model, more research needs to go into modeling the taxonomic concept circumscription—creating ontologies for morphological, genomic, or ecological traits.

2.5 Conclusions

The chapter provides an informal conceptualization of the taxonomic process and a formalization in OpenBiodiv-O. It introduces classes and properties in the domains of biodiversity publishing and biological systematics and aligns them with the important domain-specific ontologies. By bridging the ontological gap between the publishing and the biodiversity domains, it will enable the creation of Open Biodiversity Knowledge Management System, consisting of (1) the ontology itself; (2) a Linked Open Dataset (LOD) of biodiversity information (biodiversity knowledge graph); and (3) user interface components aimed at searching, browsing and discovering knowledge in big corpora of previously dispersed scholarly publications. Through the usage of taxonomic concepts, we have included mechanisms for democratization of the scholarly process and not forcing a taxonomic opinion on the users.

Chapter 3

OpenBiodiv Linked Open Dataset

We have created a dataset of biodiversity Linked Open Data, OpenBiodiv-LOD, comprising biodiversity information extracted from academic journals and public repositories. As ontology, we use the new OpenBiodiv-O developed by us (Senderov et al., 2018). We propose to the biodiversity informatics community to use OpenBiodiv-LOD as the central point for the biodiversity knowledge graph. OpenBiodiv-LOD is an RDF dataset adhering to the Principles of Linked Open Data Heath and Bizer, 2011. It is available under `http://graph.openbiodiv.net`, which provides a SPARQL endpoint for it.

OpenBiodiv-LOD is a synthetic dataset. It does not contain previously unpublished data. Instead it integrates information previously fheaound in academic journals and databases into one dataset. It also contains extracted, previously inaccessible information from the original datasets in the form of relations. In the next few sections we discuss the sources of information that were combined to from OpenBiodiv-LOD, the types of information that has been extracted, as well as the overall data model. We also discuss the Principles of Linked Open Data that tie everything together. The paper ends with many examples of queries on the dataset and with a technical discussion of how it was generated.

3.1 Dataset description

The data in OpenBiodiv-LOD comes from three major sources: from the GBIF Backbone Taxonomy (GBIF Secretariat, 2017), from journal articles published by the academic publisher Pensoft, and from Plazi Treatment Bank (`http://plazi.org`). These sources are illustrated in Fig. 3.1, which visualizes the information ow in the OpenBiodiv ecosystem. In the next subsections we describe each of these data providers in detail and the type of data that has been imported and integrated into OpenBiodiv-LOD.

3.1.1 Data from the GBIF backbone taxonomy

GBIF is the largest international repository of occurrence data. An occurrence record is a statement about the presence of an organism at a given place and time. GBIF allows its users to do searches on its occurrence data utilizing a taxonomic hierarchy. For example, it is possible to query the database for occurrences of organisms belonging to a specific genus: a search for the beetle genus Harmonia on 30 June 2018 returned 575,376 results. This search is possible thanks to the GBIF Backbone Taxonomy also known as Nub (GBIF Secretariat, 2017). Nub is a database organizing taxonomic concepts in a hierarchy covering all biological names used in occurrence records harvested by GBIF. It is a single synthetic (algorithmically generated) management classification. Thus, the GBIF backbone does not represent an expert consensus on

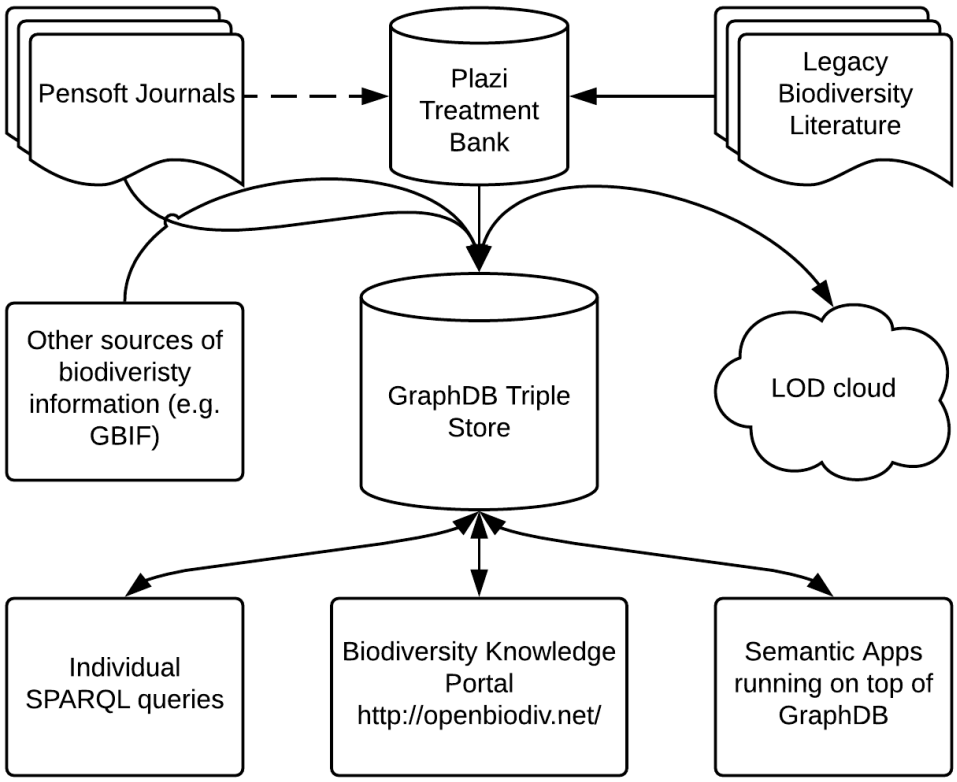


FIGURE 3.1: Sources of data and information flow in the OpenBiodiv ecosystem. OpenBiodiv-LOD is the dataset found in OpenBiodiv triple store in the center of the diagram.

how biological taxa are hierarchically arranged according to evolutionary criteria in Nature.

Keeping in mind this critique, it is evident how the backbone taxonomy allows GBIF to integrate name-based information from diverse sources of biodiversity information and to provide a facility for taxonomic searching and browsing. Some of the more known sources of information for the platform include Encyclopedia of Life (EOL), Genbank, and the International Union for Conservation of Nature (IUCN). In order to grant the same capabilities to OpenBiodiv-LOD, we have imported Nub as instances of openbiodiv:TaxonomicConcept according to OpenBiodiv-O Senderov et al., 2018. A taxonomic concept is a biological name linked to an immutable circumscription as provided by an academic publication with the help of the keyword "sec." Berendsohn, 1995. Thus, each GBIF taxonomic concept is linked to an instance of openbiodiv :ScientificName and to a resource identifying a particular version of the GBIF backbone taxonomy. Furthermore, taxonomic concepts are linked to their parent taxonomic concept via a Simple Knowledge Organization Schema (SKOS) Miles and Bechofer relation and via a fine-grained relation reified with the help of the Region Connection Calculus 5 (RCC-5) Sterner and Franz, 2017 ocabulary that OpenBiodiv-O introduces (Fig. 3.2). These links constitute the taxonomic hierarchy in the case of SKOS and, in the case of RCC-5, the network of complex inter-relations between taxonomic concepts allowing overlaps and other special cases.

The RCC-5 representation further allows the future evolution of OpenBiodiv-LOD to incorporate other simultaneous views of taxonomic alignment. For exam-ple, as the GBIF backbone taxonomy is updated regularly through an automated process from over 56 sources, future updates may be ingested as new statements into OpenBiodiv-LOD without a altering existing records: namely, as a new set of taxonomic concepts and RCC-5 relations linked to potentially already existing taxonomic names.

3.1.2 Academic journal data from Pensoft and Plazi

All valid articles from the journals published by Pensoft listed in Table 3.1 have been converted to RDF and stored in the biodiversity knowledge graph. Additionally, all valid taxonomic treatments from Plazi Treatment Bank have been converted to RDF and stored in the graph as well. A taxonomic treatment is the special part of a taxonomic publication where the taxonomic concept circumscription (species description) takes place. Furthermore, the RDF-ization procedure is triggered automatically on a weekly basis and thus the semantic database is always updated with the newest articles published by Pensoft and newest taxonomic treatments extracted by Plazi. The RDF-ization is made possible by the fact that all Pensoft journals are published as XML according to TaxPub, an extension of the NLM/NCBI journal publishing DTD for taxonomic description (Catapano, 2010)and, similarly, all Plazi treatments follow the TaxonX XML Schema (Penev et al., 2011) Thus, the RDF-ization pipeline does not require a natural language processing step, as a considerable amount of information is marked-up at the time of publication. We have given an example of how a taxonomic name usage is marked up in a TaxPub article in Listing 7.1. The datatypes that have been marked up in TaxPub and TaxonX and whose entities are converted to RDF and integrated in OpenBiodiv-LOD are listed in Table 3.2Note that the marked-up datatypes do not correspond one-to-one to the RDF entities that have been created in the graph as TaxPub, TaxonX, and OpenBiodiv-O take slightly different approaches to modeling the biodiversity world. OpenBiodiv-O takes the most granular approach.

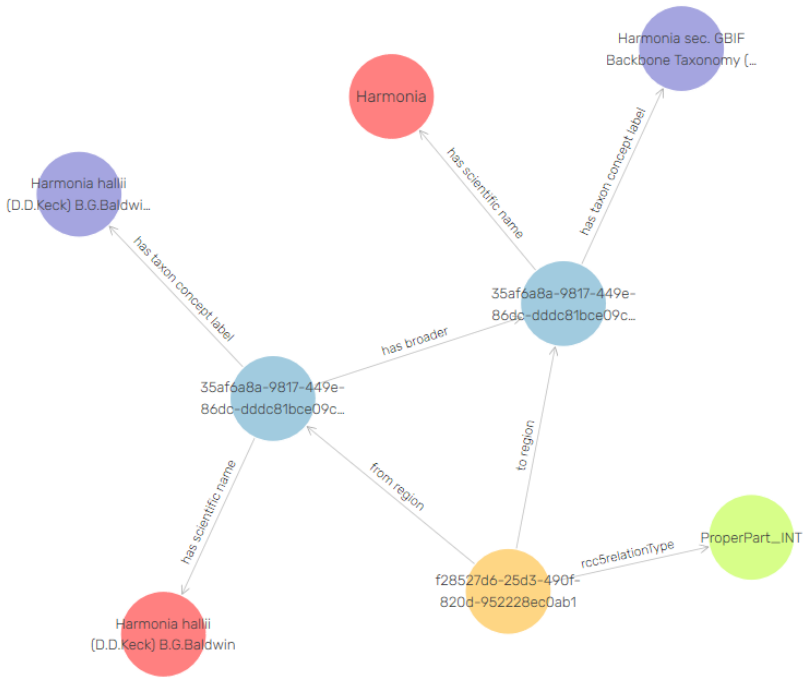


FIGURE 3.2: Illustration of the representation of hierarchical information imported from the GBIF Backbone Taxonomy as two taxonomic concepts, *Harmonia halii* sec. GBIF Secretariat, 2017 and *Harmonia* sec. GBIF Secretariat, 2017. Each concept has an associated scientific name via *has scientific name*; however, the hierarchical information is not encoded in the names. The hierarchical relationship between *Harmonia halii* sec. GBIF Secretariat, 2017 and *Harmonia* sec. GBIF Secretariat, 2017 is encoded both as SKOS *has broader* and reified via the RCC-5 relationship encoded in `f28527d6-25d3-490f-820d-952228ec0ab1`.

For example, each taxonomic name usage in a Pensoft article results in a corresponding `openbiodiv:TaxonomicNameUsage` resource and a link to the `openbiodiv:ScientificName` resource that the taxonomic name usage mentions (Fig. 3.3).

TABLE 3.1: RDF-ized biodiversity journals published by Pensoft.

Journal Name	Submission Style	Number of Articles
ZooKeys	Word document	3829
PhytoKeys	Word document	537
MycoKeys	Word document	127
Biodiversity Data Journal	Web based (ARPHA)	490
Journal of Orthoptera Research	Word document	32

TABLE 3.2: Datatypes marked up in TaxPub and TaxonX articles and the corresponding RDF types of the generated RDF resources. The TaxPub and TaxonX columns contain boolean values indicating whether the information about the datatype is retrieved from files encoded in the corresponding schema.

Datatype	TaxPub	TaxonX	RDF Type
Article metadata	T	T	`fabio:JournalArticle` and related
Keyword group	T	F	`openbiodiv:KeywordGroup`
Abstract	T	T	`sro:Abstract`
Title	T	F	`doco:Title`
Author	T	T	`foaf:Person`
Introduction section	T	F	`deo:Introduction`
Discussion section	T	T	`orb:Discussion`
Treatment section	T	T	`openbiodiv:Treatment`
Nomenclature section	T	T	`openbiodiv:NomenclatureSection`
Materials examined	T	T	`openbiodiv:MaterialsExamined`
Diagnosis section	T	T	`openbiodiv:DiagnosisSection`
Distribution section	T	T	`openbiodiv:DistributionSection`
Taxonomic key	T	T	`openbiodiv:TaxonomicKey`
Figure	T	T	`doco:Figure`
Taxonomic name usage	T	T	`openbiodiv:TaxonomicNameUsage`

3.2 Example of SPARQL queries

We shall illustrate and evaluate the LOD by issuing sample SPARQL queries illuminating aspects of it.

3.2.1 Simple queries

Query for author

Authors are instances of `foaf:Person` (except in the rare institutional case, in which case they would be `foaf:Agent`). The SPARQL query in Listing 7.2 answers the question of which authors have been the most prolific.

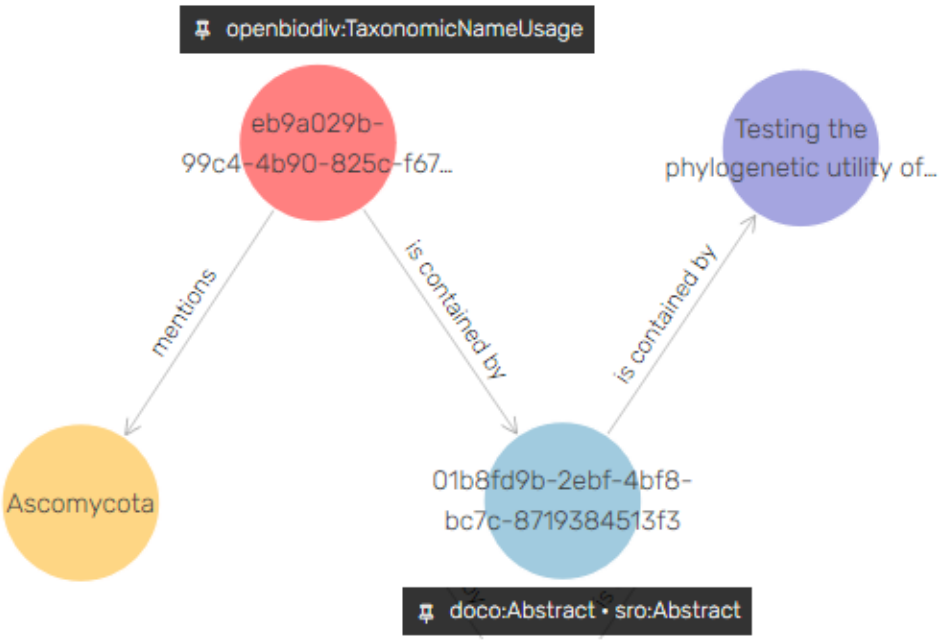


FIGURE 3.3: The taxonomic name usage (`openbiodiv:eb9a029b-99c4-4b90-825c-f670fb88900d`) is linked to the scientific name it mentions, *Ascomycota* and to the part of the article (abstract) that it is contained in.

Query for scientific name

Latin names are stored in the system as `:ScientificName` and are mentioned by taxonomic name usages. Listing 7.3 orders scientific names of any rank by the number of unique mentions that they have in articles. It is possible to narrow down the solution to binomial names (species names) by adding the `dwciri:taxonRank` property as shown in Listing 7.4. In order to see the different ranks that scientific c names are assigned to in OpenBiodiv we can use the query in Listing 7.5. It is also possible, for example, to determine the most mentioned scientific c name by the number of articles it is mentioned in Listing 7.6.

Query the article structure

A unique feature of OpenBiodiv-LOD is that articles are broken down into their components (see e.g. Table 3.2) and taxonomic name usages are connected to the specific part of the article and not just to the article in general. Combining this feature with queries from the previous paragraph, we can, for example, look for the most mentioned scientific name in a figure (Listing 7.7), or for the figures present in a particular article (Listing 7.8).

Query for taxonomic concepts

We can create a query uniting information from concepts from the GBIF Backbone Taxonomy with semantics coming from the article structure. The query in Listing 7.9 locates taxa that are in the beetle family Curculionidae according to taxonomic backbone of GBIF (sec. GBIF Secretariat, 2017) in the system, and looks for new taxa

(:TaxonomicDiscovery)and looks for new taxa (:TaxonomicDiscovery) that have been associated with one of its genera.

Fuzzy Queries via Lucene

The SPARQL endpoint of OpenBiodiv-LOD sup-ports fuzzy matching via a Lucene connector (Ontotext, 2018). This can be a very useful as due to multiplicity of taxonomic names and the complexities of Latin grammar, one often does not remember the correct spelling of a name. The Lucene query needs to follow the standard Lucene query syntax (The Apache Software Foundation, 2013) and is specified as a literal string of the property `http://www.ontotext.com/connectors/lucene#query` of the search variable Listing 7.10.

3.2.2 Competency question answering via SPARQL

At the end of Chapter 2 we suggested some competency questions that may be answered by OpenBiodiv.

Validity of a taxonomic name

Of central importance is the question of whether a given taxonomic name is valid or not. We shall consider a taxonomic name invalid if and only if at least one of the following invalidation criteria holds:

1. The name has been replaced. I.e. there is a `:replacementName` property originating in the name and there are no loops: it is impossible to follow the `:replacementName` edges and come back to the name. This query is illustrated in Listing 7.11.

2. The name has been invalidated, i.e. there is a taxonomic usage with the status `:UnavailableName` and there is no newer taxonomic name usage revalidating it (`:AvailableName`). Illustrated in Listing 7.12.

Investigation of the impact of the lost collections of Museu Nacional

In order to illustrate the capabilities of OpenBiodiv and draw attention to the impact of the tragically lost collection of the Museu Nacional de Rio de Janeiro (MNRJ), we can ask our system to give us the number of times a specimen from that collection was used in a taxonomic article, and in which ones (Listing 7.13). It turns out that MNRJ has been mentioned 195 times in our system in a total of 22 articles published by Pensoft. Perhaps more interestingly, we can see specimens of which taxa may have been lost. Examples include the insects (Xestoblatta, Charinus, Lamproclasiopa, etc.), nematode worms (Paracamallanus, Cucullanus, Pseudascarophis, etc.), birds (Ichthyouris), sh (Sphoeroides), and many others for a total of 6,127 distinct names mentioned in taxonomic articles whose materials methods include MNRJ.

3.3 Methods

The inputs are either XML (Pensoft and Plazi) or CSV (GBIF). Thus, the raw datastreams are semi-structured and the dataset generation problem can be thought of as an information retrieval and transformation problem. The input is encoded in three different data models|DarwinCore CSV (GBIF), TaxPub XML (Pensoft), and TaxonX XML (Plazi). The output of the transformation pipeline is knowledge represented in a fully-structured RDF according to the ontology.

3.3.1 Obtaining the data

The first step before running any transformation is to obtain the raw inputs. GBIF's taxonomic backbone is available under
https://www.gbif.org/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c.
There is an RSS feed from which Plazi's treatments can be downloaded on a daily basis under http://tb.plazi.org/GgServer/xml.rss.xml. Each of Pensoft's journals has a public API endpoint under `http://[journal_name].pensoft.net/lib/journal_archive.php`, where `[journal_name]` ought to be replaced with the name of the Pensoft journal. E.g. `bdj` to make http://bdj.pensoft.net/lib/journal_archive.php.

3.3.2 Tools

In order to carry out the dataset generation we made use of the following tools:

1. RDF4R R package[1], which is described in Chapter 4 and deals with all RDF-related issues such as accessing a triple store, serializing the in-memory resource representations to Turtle files, etc.

2. `ropenbio` R package[2], which implements the data retrieval and transformations described in this chapter.

3. TSV4RDF, which is a PHP library for mapping CSV to RDF developed by Pensoft. It is developed outside of the scope of the dissertation and is not discussed in detail.

4. The OpenBiodiv base package[3], which contains scripts needed for the initialization and updating of the database.

In the rest of the section we describe the transformation from XML as it is implemented in ROpenBio.

3.3.3 XML to RDF transformation

In order to transform an article represented as an XML document to RDF, we make use of the hierarchical nature of XML and solve the problem recursively with the following Extractor procedure in Algorithm 1. The extractor's procedure input is an XML node and its output is the RDF corresponding to the XML node. The extractor procedure has three essential steps: atoms extraction, RDF constructions from the extracted atoms, a divide-and-conquer step that recursively calls itself and unites the results. Extraction of a whole article is achieved by calling the Extractor on the root node of the article.

Atoms extraction

The atoms of an XML node consist of all text-fields that can be reached from the XML node with an XPATH expression (can be attribute values or text values) that can be directly converted to RDF as literals or identifiers. They all belong to one or to several related RDF resources. For example in Listing 7.14 we have listed the XML node that

[1] RDF4R package on GitHub: github.com/pensoft/rdf4r
[2] ROpenBio R package on GitHub: github.com/pensoft/ropenbio
[3] OpenBiodiv documentation and scripts: github.com/pensoft/OpenBiodiv

Algorithm 1 The Extractor procedure

```
1: procedure EXTRACTOR(XML Node X)
2:     a ← extract atoms of X                          ▷ Atoms extraction
3:     r ← construct RDF from a                        ▷ RDF construction
4:     C ← find relevant sub-nodes of X                ▷ Recursively applies itself
5:     R ← apply Extractor on each C_i ∈ C
6:     return r ⋃ R
7: end procedure
```

contains author information in the TaxPub schema. The atoms here are `surname = "Wachkoo"`, `given_name = "Aijaz Ahmad"`, `orcid_id = "https://orcid.org/0000-0003-2506-9840"`, `affiliation = "Central Institute of Temperate Horticulture, Srinagar, Jammu &Kashmir, India"`. In order to achieve the extraction, the atoms extractor must know the XPATH locations (e.g. the surname is at `./name/surname`) of the authors it is looking for and the types of the values (e.g. string, integer, link, etc.). Sometimes this can be quite challenging as is the affiliation field in the given example. In it the XPATH location of the address string is influenced by the value of `xref`. We were, however, unable to express this situation in pure XPATH. For example `//aff[id=./xref/@rid]` is the wrong idea: here . does not refer any more to the author object object but rather to the last matched object, i.e. the `aff` object.

RDF Generation

Once the atoms have been extracted they can be put together as RDF. Conceptually this is straightforward as for each atom we know its type and therefore we know which RDF property to use. The author example is given in Listing 7.15.

It should be noted in this paragraph that the semantics of certain node types such as taxonomic name usage (reified as `:TaxonomicNameUsage`) reflect the relative position of the node in the XML document. For example, a taxonomic name usage may be inside a figure, inside an introduction section, inside a title, etc. Therefore besides the atoms, the constructor receives information about the relative position of the resource in the article by means of the unique identifier of the parent node(s). Then this information is encoded in RDF as given in Listing 7.16.

Divide and conquer

After we have successfully converted the current XML node to RDF, a recursive call to Extractor is made for all nodes that are hierarchically dependent on the current node. For example, the article node contains all the other other nodes such as sections, figures, etc.

Transformation specification

In order for the Extractor to work, therefore, we need to specify an XML schema. The specification includes what XML nodes we are looking for and their location. It then recursively specifies for each node, what sub-nodes we are looking for and their XPATH location relative to their parent node. Finally, for every node we need to give the atom locations and write a constructor. The transformation specification is

done with R6 framework in R. We have specified two schemata that share the same constructors—TaxPub[4] and TaxonX[5].

3.3.4 Submission to graph database and post-processing

In the previous section we described how we transform XML documents in TaxPub and TaxonX to RDF statements according to OpenBiodiv-O. In addition, we transform the GBIF backbone taxonomy to RDF according to OpenBiodiv-O with the help of TSV4RDF, a proprietary Pensoft tool. The generated RDF statements are submitted to a repository in a GraphDB instance residing on `http://graph.openbiodiv.net/`. The repository has been initialized with OpenBiodiv-O and the ontologies on which it depends[6]. Finally, after the data has been submitted, update scripts are run to generate further statements from our ontology that have not been encoded in OWL for the updating of scientific name relations.

Update rule for replacement name

We state that a scientific name A replaces a scientific name B, if there exists a taxonomic name usage of A with taxonomic status `:ReplacementName` and B is mentioned by a taxonomic name usage in the nomenclatural citations of the treatment, where the discussed taxonomic name usage of A is in the nomenclature section (Listing 7.20).

Update rule for related name

The related names update-rule is similar to the replacement name: two scientific names A and B are considered related if they both mentioned in the nomenclature section of a treatment (Listing 7.18).

3.4 Discussion

3.4.1 Fulfillment of the Principles of Linked Open Data

Linked Open Data (LOD, Heath and Bizer, 2011) is an idea of the Semantic Web (Berners-Lee et al., 2001) aimed at ensuring that data published on the Web is reusable, discoverable and, most importantly, that pieces of data published by different entities can work together. The principles of LOD are the following (Heath and Bizer, 2011)

1. Use URIs as names for things.
2. Use HTTP URIs so people can lookup these things.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs so they can discover more things.

We have followed these guidelines when creating the OpenBiodiv LOD. We will now discuss each of these points separately.

[4] `https://github.com/pensoft/ropenbio/blob/redesign/R/taxpub.R`
[5] `https://github.com/pensoft/ropenbio/blob/redesign/R/taxonx.R`
[6] `https://github.com/pensoft/openbiodiv-o/tree/master/imports`

Usage of URIs as resource identifiers

Every instance in OpenBiodiv LOD is uniquely identifiable by a HTTP URI of the following form: `http://openbiodiv.net/uuid-(suffix)`. All instance identifiers in OpenBiodiv LOD follow this schema. The optional suffix field is assigned only to resources extracted from GBIF.

Identifiers for Pensoft and Plazi. During the RDF-ization of the sources Pensoft and Plazi, when a new concept is discovered (e.g. a person, a scientific name, etc.) a UUID is generated. Then the resource is always referred to in the database by this UUID in the OpenBiodiv namespace, http://openbiodiv.net/. Pensoft and Plazi furthermore share the UUID part of the identifier in the semi-structured representation of treatments. For example, Lyubomir Penev is a resource identified by http://openbiodiv.net/7a247614-878b-4d01-ab97-bf5fc608dc86.

Identifiers for GBIF taxonomic concepts. GBIF offers its taxonomic backbone as a big DarwinCore (Wieczorek et al., 2012) tab separated file (TSV). Each row in the TSV corresponds to a taxonomic concept published by GBIF. GBIF does not offer a globally unique ID of its concepts, but only a local ID (e.g. 4239 is the GBIF ID of concept of Curculionidae sec GBIF Secretariat, 2017). This is why, we generated a UUID (e.g. for the example `35af6a8a-9817-449e-86dc-dddc81bce09c-4239`) for each row of data table published from GBIF. However, each taxonomic concept is linked to a taxonomic name and to a taxonomic concept label (see Chapter 2). It was impractical for programmatic reasons to generate a new UUID for these linked entities. This is why their unique identifiers are suffixed. We use the suffix `-ScientificName` to denote scientific names and `-TCL` to denote taxonomic concept labels.

In our example we have respectively `http://openbiodiv.net/35af6a8a-9817-449e-86dc-dddc81bce09c-4239-ScientificName` and `http://openbiodiv.net/35af6a8a-9817-449e-86dc-dddc81bce09c-4239-TCL`.

Usage of HTTP URIs and dereferencing

As per the Linked Data Principles, we use dereferenceable HTTP URIs for our resources. For example, if a web-browser opens `http://openbiodiv.net/35af6a8a-9817-449e-86dc-dddc81bce09c-4239-ScientificName` a web-page is displayed (Fig. 3.4) providing useful information for the name such as where it used and other names are related to it. Also it is possible to request OpenBiodiv resources via Curl with the header `Content-Type: application/rdf+xml` and an RDF representation of the resources is returned.

Linking to other resources

First, all resources in OpenBiodiv form a graph (there are no disconnected parts). The data model is discussed in the next section. Second, taxonomic names are linked to external databases via `dwc:taxonID`. These are strings containing GBIF ID's, ZooBank ID's, LSID's, etc. Unfortunately as HTTP URI's have not gained popularity in the biodiversity informatics community, the only true resource-id-to-resource-id links are within OpenBiodiv itself. However, we hope that the introduction of OpenBiodiv LOD contributes to the amelioration of this situation.

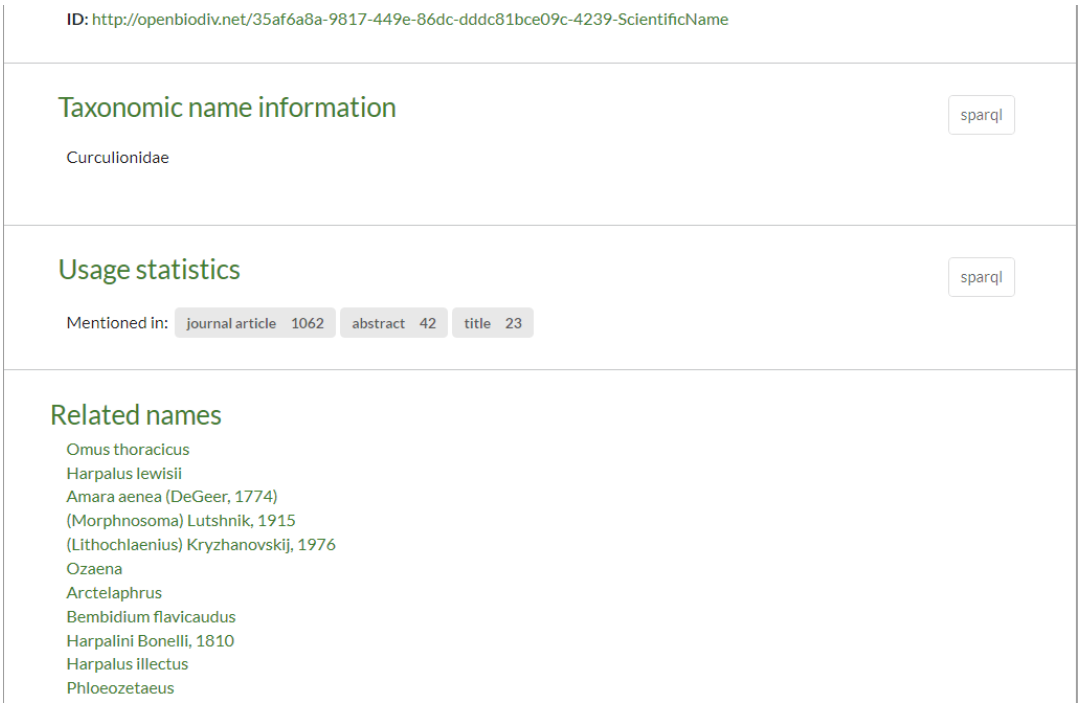


FIGURE 3.4: Visualization.

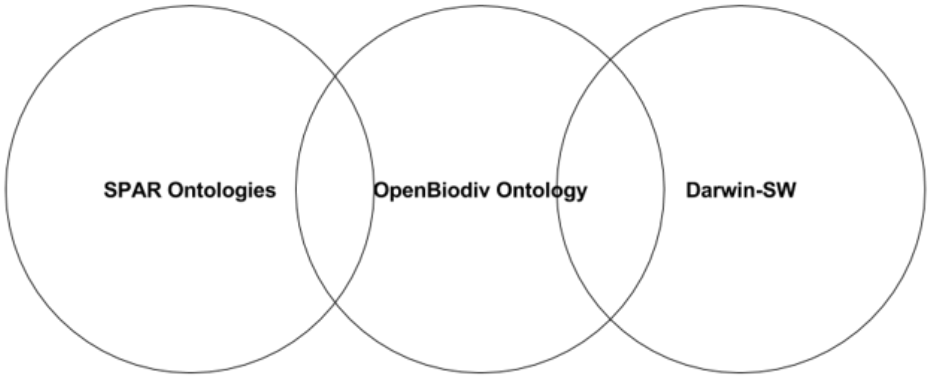
3.4.2 Data Model

When creating the RDF graph we have conformed to the OpenBiodiv Ontology described in Chapter 2 and well-established community ontologies (Fig. 3.5). In particular, (1) we use the Semantic Publishing and Referencing Ontologies (SPAR, Peroni, 2014) to model entities from publishing such as Journal, Article, Section, Figure, Table, and so on; and (2) we use the DarwinCore (DwC, Wieczorek et al., 2012) community standard and its extension, the Darwin-SW (Baskauf and Webb, 2016) ontology, to model entities the biodiversity domain.

SPAR provides facilities to deal with the dichotomy between the abstract representation of knowledge through the class Work and its concrete representation through the class Expression. For example, a `fabio:JournalArticle` can be the realization of a `fabio:ResearchPaper`. On the other hand, the DwC community standard gives a standard way to express properties from taxonomy and biodiversity science and its extension Darwin-SW a way to reify elements of an occurrence instance such as Identification, Organism, Token, and so on. A caveat: the current version of OpenBiodiv-LOD does not store yet occurrence information but all necessary infrastructure is in place to include them in the next release.

3.4.3 Observation on performance

The current iteration of the database holds over 600 million triples (Fig. 3.6). The expansion ratio under the RDFS-Plus (Optimized) ruleset is 2.35, i.e. for each asserted statements we materialize on average 2.35 implicit statements. Under the OWL2-RL ruleset (which contains a full implementation of the OWL logic), the expansion ratio is about 3.7; however, we encountered significant performance issues using it (Fig. 3.7). Even with the lighter ruleset (RDFS-Plus Optimized), we still see performance degradation with increasing database size. Importing the GBIF backbone taxonomy from



Work	Treatment	Identificiation
Expression	Nomenclature	Organism
Journal	Nomenclature Heading	Token
Journal Article	Taxonomic Name	Occurrence
Research Paper	Latin Name	Event
FrontMatter	Taxonomic Concept Label	Agent
Body Matter	Taxonomic Concept	Location
Back Matter	Taxonomic Status	
Section	Vocabulary	
Table	RCC5 Vocabulary	
Figure		

FIGURE 3.5: OpenBiodiv-O is an ontology that links the publishing domain with the biodiversity domain. Major resource types covered by each of the ontology families are given in the box below the Venn diagram. Important resources from the publishing domain are listed in the leftmost column and from biodiversity informatics in the rightmost column. The middle one covers important OpenBiodiv-O resources.

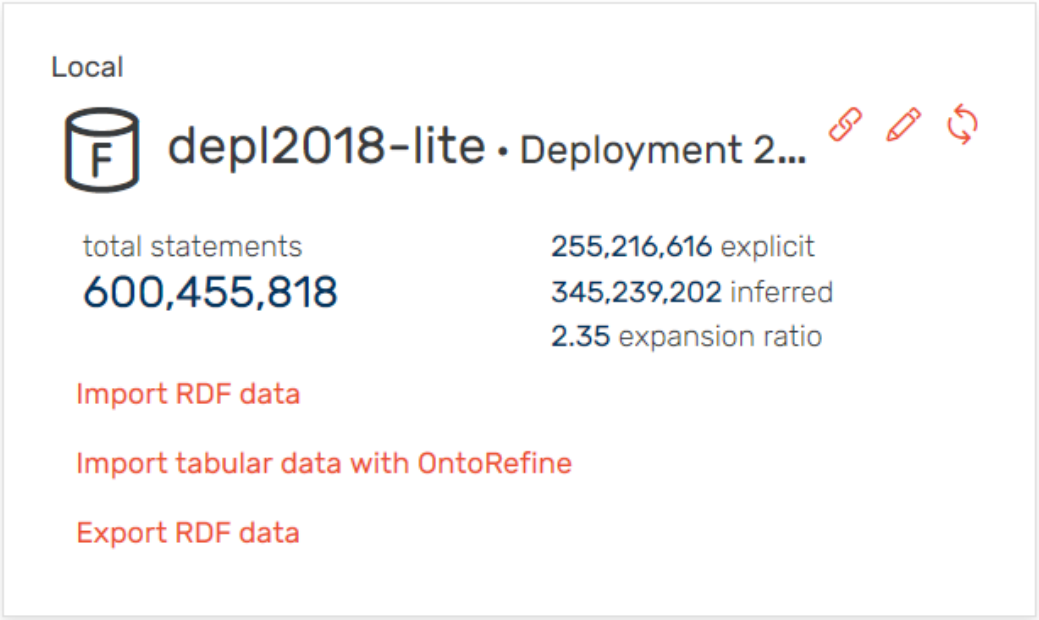


FIGURE 3.6: Statements report from the GraphDB workbench.

file takes about two days under the easier scenario. The subsequent importing of the Pensoft archives takes about two weeks as it is a slower operation requiring not only the time for submission but the time for converting the XML's to RDF.

3.5 Conclusion

The generated dataset OpenBiodiv-LOD, similar to the ontology OpenBiodiv-O, is already a solid resource for biologists, as it includes information from most articles published by Pensoft and Plazi and counts over 600 million RDF triples. Like the ontology, it ought to be further expanded.

An important conclusion that can be drawn from the work is that it is possible to use a semantic graph for the integration of a large volume of data on biodiversity. We were unexpectedly given the opportunity to illustrate the power of the knowledge graph by analyzing the damage from the tragic re at the Museu Nacional in Rio de Janeiro. In addition, we have illustrated that it is possible to write relatively simple logical conclusions to check the validity of a taxonomic name.

Due to the large amount of data, we found that although the use of a semantic graph was possible, some of the initially chosen technologies proved to be inapplicable or difficult to apply. We have observed that the practical application of the full logical OWL model is difficult due to performance problems. Instead in the end, we utilized RDFS that is less powerful but faster.

A big difficulty was the disambiguation of resources such as author names or taxonomic names. In the functional design of the RDF4R package we have put modules that allow us to insert a list of functions/rules for disambiguation when searching for an identifier for a given resource. However, we had only limited success with the rule-based disambiguation and for this reason in the production system it was discontinued at the moment.

Considering these and other "lessons," the future development of the OpenBiodiv project can be outlined in the following not necessarily comprehensive way:

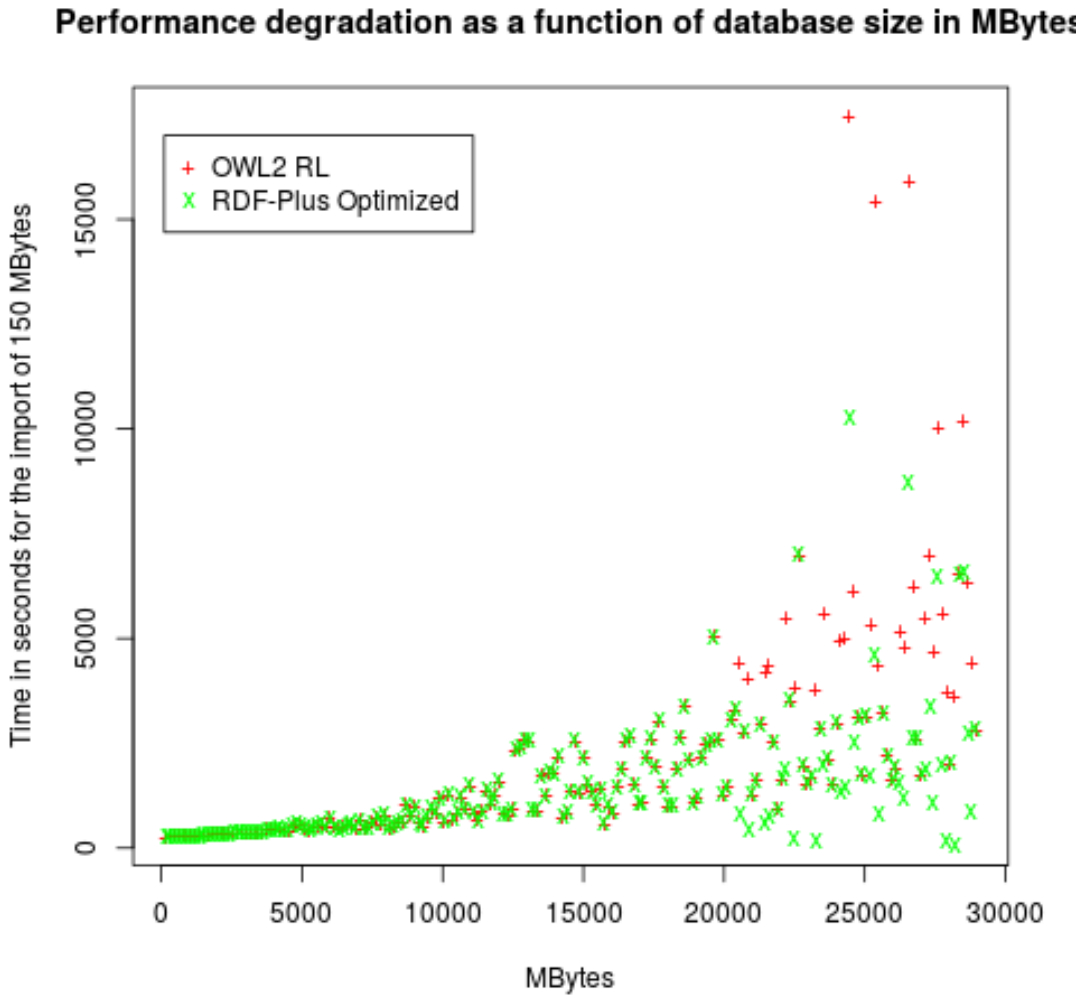


FIGURE 3.7: The graph visualizes the time in seconds needed to import a 150 MB big Turtle data file as a function of the database size. The database size is measured by the adding up the size of the data files that have already been imported.

1. As an immediate goal, to expand the LOD and ontology with new data types and new data sources using the existing framework. Such data are e.g. genomic data, occurrence data, (bio-)geographic data, visual data, descriptive data, etc.

2. Look for even closer integration with other existing biodiversity data repositories than GBIF. For example, BioImages, iNaturalist, BOLD, and so on.

3. As a longer-term task to study the transition from a semantic graph to a technology where the inference engine is separated from the data base layer as WikiData or Neo4j. In addition to increased performance, this will give extra flexibility to the project, such as allowing the use of non-RDF-based inference engines such as Euler.

Chapter 4

RDF4R: R Library for Working with RDF

RDF4R (`rdf4r`) is an R package for working with Resource Description Framework (RDF Working Group, 2014) data. It was developed as part of the OpenBiodiv project but is completely free of any OpenBiodiv-specific code and can be used for generic purposes requiring tools to work with RDF data in the R programming environment (R Core Team, 2016).

4.1 Prerequisites

RDF4R depends on the following packages (list may change in future releases):

- `gsubfn` (Grothendieck, 2018)
- `httr` (Wickham, 2017)
- `xml2` (Wickham et al., 2018b)
- `R6` (Chang, 2017)
- `devtools` (Wickham et al., 2018a)—needed if one is to do a GitHub install.

We are currently in the process of submitting the package to a repository and making it available through the standard installation facilities of R. Our intention is to publish it through CRAN and/or rOpenSci[1].

4.2 Specification

RDF4R has the features listed in the following subsections.

4.2.1 Connection to a triple-store

Triple-stores, also known as quad-stores, graph databases, or semantic databases, are databases that store RDF data and allow the querying of RDF data via the SPARQL query language (The W3C SPARQL Working Group, 2013). RDF4R can connect to triple-stores that support the RDF4J server REST API (RDF4J development team, 2017) such as GraphDB (Ontotext, 2018). It is possible to establish both basic connections (requiring no password or requiring basic HTTP user-pass authentication) or connection secured with an API access token.

[1]Repositories accessible under `https://cran.r-project.org/` and `https://github.com/ropensci/`, respectively.

4.2.2 Work with repositories on a triple-store

Once a connection to a triple-store has been established, it is possible to inspect the talk protocol version, view the list of repositories on the database, execute SPARQL Read (SELECT keyword and related) and SPARQL Update (INSERT and related) queries on the database, as well as submit serialized RDF data directly to the database.

4.2.3 Function factories to convert SPARQL queries to R functions

An important feature of RDF4R are its facilities for converting SPARQL queries and the like to R functions. This conversion is realized by a family of functions that return functions. In this thesis they will be referred to as function factories. Taking the example of the `query_factory` we have: given a parameterized SPARQL query (parametrization syntax is explained later), `query_factory` returns a function f whose arguments are the parameters of the query. Upon being called f submits the query to a SPARQL endpoint and returns the results.

4.2.4 Work with literals and identifiers

The building blocks of RDF are literals (e.g. strings, numbers, dates, etc.) and resource identifiers. RDF4R provides classes for literals and resource identifiers that are tightly integrated with the other facilities of the package.

4.2.5 Prefix management

Prefixes are managed automatically during serialization by being extracted from the resource identifiers.

4.2.6 Creation and serialization of RDF

RDF4R uses an own implementation of smart dynamically reallocated vector data structure to store RDF triples as mutable R6 objects, called `DynVector`. Blank nodes are partially supported: a triple may contain an anonymous RDF object (a list of triples with the same subject) as its third position. In this case, the parent RDF is serialized as Turtle by using the bracket syntax, i.e. to express that someone knows someone else whose name is "Bob", we can write in Turtle `:someone foaf:knows [ foaf :name "Bob"] .` Currently, the serialization procedure only supports Turtle (and its variant Trig, Bizer and Cyganiak, 2014) and only supports adding new triples.

4.2.7 A basic vocabulary of semantic elements

RDF4R has some basic resource identifiers for widely used classes and predicates predefined (e.g. for `rdf:type`, `rdfs:label`, etc.).

4.3 Usage

Here, we explain how to use the package RDF4R by means of examples. In order to fully utilize the package capabilities, one needs to have access to an RDF graph database. We have made available a public endpoint (see next paragraph) to allow the users of the package to experiment. Since write access is enabled, please be considerate and don't issue catastrophic commands.

4.3.1 Connection to triple store

The code in Listing 7.19 creates an object, `graphdb`, that stores the information needed to access the database. The user needs to supply it to the access functions discussed later. We have also created an object `openbiodiv` that contains read-only access to the main OpenBiodiv instance. All examples in this section use one of these two access points.

4.3.2 Example: convert a SPARQL query to an R function

The purpose of this example (Listing 7.20) is to convert a simple SPARQL lookup query to an R function. The publicly accessible endpoint happens to store some biological information but for the purposes of this example knowledge of biological taxonomy or the ontology of the store is irrelevant. It is only necessary to know that the database stores information about biological papers that contain references to biological names. We are looking for papers that mention the biological name *Drosophila*, which is a genus of flies. Please note the parametrization of the SPARQL query.

Note that this almost valid SPARQL with one exception: the `%param` string on the first line of the WHERE clause. A SPARQL query is parameterized by specifying a % in front of the tokens that have to become the arguments of the generated R function. In order to construct a function that looks up a genus (a biological rank) by a supplied string we then simply write: `genus_lookup = rdf4r::query_factory(p_query = p_query, access_options = openbiodiv)`.

`query_factory` takes two arguments: the parameterized query and an object with access options for an endpoint and returns an R function whose arguments are the parameters of the parameterized SPARQL query and which executes the SPARQL query against the endpoint specified in `access_options` and returns formatted results as a data frame.

The usage of the function that we have just created is given in Listing 7.2. Note that we have enclosed the string `Drosophila` in escaped quotes as only that would make the replacement of the parameter in the parameterized SPARQL query a valid SPARQL. Had the parameter been a resource identifier, we would not have needed the quotes. In a later example, we will show how we can get around this hassle by utilizing the built-in classes for literals and resource identifiers.

Excercise. Try experimenting with `genus_lookup` by looking up information about some other genera (*Eupolybothrus*: a millepede, *Myotis*: a bat).

4.3.3 Setting up literals and identifiers

We want to model Table 4.1 (recreated from Alemang and Hendler 2008). We will use the prefix `<http://rdflib-rdf4r.net/>` for all instances that we create and a dummy ontology (not actually defined) with the prefix `<http://art-ontology.net/>` to reify the example classes and properties.

Literals

In Listing 7.21, we repeatedly called the `literal` function, which is a constructor of objects of class *literal*, with different arguments. `literal` can construct phrases in English (or any other language) with the `lang` argument. It can construct pure strings (by omitting the `lang` argument) of type `xsd:string`. It can also construct literals

TABLE 4.1: Tabular Data about Elizabethan Literature and Music recreated from Allemang and Hendler, 2011.

ID	Title	Author	Medium	Year
1	*As You Like It*	Shakespeare	Play	1599
2	*Hamlet*	Shakespeare	Play	1604
3	*Othello*	Shakespeare	Play	1603
4	"Sonnet 78"	Shakespeare	Poem	1609
5	*Astrophil and Stella*	Sir Philip Sidney	Poem	1590
6	*Edward II*	Christopher Marlowe	Play	1590
7	*Hero and Leander*	Christopher Marlowe	Poem	1593
8	Greensleeves	Henry VIII Rex	Song	1525

of a number of defined semantic types by using the `xsd_type` argument. In order to see which types are available execute `?semantic_elements`. Note that the XSD types are implemented as resource identifiers (class *identifier*), which allows the user to implement additional types that are not provided. Behind the scenes the URL of the resource identifier will be appended to the representation (`?represent`) of the literal as text through pasting of `\textasciicircum\textasciicircum` and then the URL.

In other words, for the work titles, we use the argument `lang="en"` telling the literal constructor that the literal value is in English, whereas for the names, we omit this argument. As per semantic web conventions, when the argument is omitted, and no type is explicitly specified, it is assumed that the literal is a string (`xsd:string`). For the literals containing years, on the other hand, we explicitly specify an integer type; otherwise they would have parsed as strings as well. All of this can be seen by inspecting the individual lists in Listing 7.22 (objects of class *literal* are lists).

Identifiers

We need resource identifiers for our resources, i.e. playwrights, works of art, as well as for the classes of which those resources are instances of. To make things simpler, we use a fictional ontology with the prefix `http://art-ontology.net/`. We hard-code identifiers for the ontology classes (Listing 7.23, 7.24).

Note that each identifier object is a list where the field `$uri` gives the URI of the resource and the field `$qname` gives the shortened name (QNAME) with respect to the prefix stored in `$prefix`. Also note that both literal and identifier are *representable*, i.e. we have defined a the generic represent on both of the classes that outputs a proper string representation of the literal or resource identifier that can be used in a serialization.

We also need resource identifiers for our entities such as Shakespeare, Christopher Marlowe, etc. Semantic Web best practices discourage the liberal minting of identifiers for resources for which somebody has already minted an identifier. Instead, we want to look them up in a database, and only mint if they are not found. For this, RDF4R offers factory functions to create lookup/ mint functions as seen in Listing 7.25.

`identifier_factory`'s first argument, `fun`, is a list of (lookup) functions that will be tried. `identifier_factory` returns an identifier constructor function, in our case we named it `lookup_or_mind_id`. The lookup functions need to return a single column (labeled e.g. `?id`). They will be tried in order and if any of them returns a unique solution, it will be returned by the constructor function to create an identifier object. If none of the lookup functions returns a unique solution, a new identifier will be minted.

4.3.4 Creating RDF

To create an RDF representation, create a new `ResourceDescriptionFramework` object and add triples to it (Listing 7.26).

The easiest way to inspect the `ResourceDescriptionFramework` object is to serialize it. Before we serialize, however, we need to specify the subgraph where the triples should be stored with `$set_context(id)`. We will reuse the example for that (Listing 7.27).

4.3.5 Submitting RDF to the triple store

Now that we have created some RDF (`classics_rdf`), we are ready to submit it to the endpoint (`graphdb`). We can either submit it directly via `add_data`, or we can use the `add_data_factory` to create a submitter function (Listing 7.28).

4.4 Discussion and Conclusions

4.4.1 Related Packages

The closest match to RDF4R is the `rdflib` (Boettiger, 2018). The development of the two packages was simultaneous and independent until `rdflib`'s first official release on Dec 10, 2017. This explains why two closely related R packages for working with RDF exist. After the release of `rdflib` work was started to make both packages compatible with each other. In our opinion, the packages have different design philosophies and are thus complementary.

`rdflib` is a high-level wrapper to `redland` (Jones et al., 2016), which is a low-level wrapper to the C `librdf` (Beckett, 2014), a powerful C library that provides support for RDF. `librdf` provides an in-memory storage model for RDF beyond what is available in RDF4R and also persistent storage working with a number of databases. It enables the user to query RDF objects with SPARQL. Thus, `librdf` can be considered a complete graph database implementation in C.

In our opinion, `redland` is more complex than needed for the purposes of OpenBiodiv. By the onset of the OpenBiodiv project it was available[2]; however, we decided not to use it as a decision was made to rely on GraphDB for our storage and querying. Note that RDF4R's main purpose is to provide a convenient R interface for users of GraphDB and similar RDF4J compatible graph databases.

A feature that differentiates `rdflib` from RDF4R is the design philosophy. RDF4R was designed primarily with the Turtle and TriG serializations in mind. This means that RDF4R can work with named graphs, whereas their usage is discouraged or perhaps impossible with `rdflib`[3], even though `rdflib`'s default format is N-Quads.

Another differentiating feature between RDF4R and `rdflib` is that RDF4R provides facilities for converting SPARQL and related statements to native R functions!

In a future release of RDF4R (2.0) we would like to replace or extend its in-memory model with `rdflib`'s. This is why we would like to make the packages fully compatible and have contributed several patches to `rdflib`[4]). Thus, it will be possible for the user of RDF4R to retain its syntax and high-level features— constructor factories, functors, etc., and the ability to use named graphs—but benefit from performance increases, stability, and scalability with the `redland/rdflib/librdf` backend.

[2] But not `rdflib`!

[3] The issue was discussed on the `librdf` GitHub page, https://github.com/ropensci/rdflib/issues/23.

[4] Please, consult the commit history under https://github.com/ropensci/rdflib.

This will enable the users of the R programming environment to use whichever syntax they prefer and benefit from an efficient storage engine.

4.4.2 Elements of Functional Programming (FP)

When choosing a programming environment for the project, a choice was made to use R (R Core Team, 2016) given its ubiquity in data science. Having settled on R, we decided to incorporate elements of the functional programming style (Wickham, 2015) that R supports in order to allow the users of the package to write simple code to solve complex tasks.

RDF4R is not written as in a pure functional programming style: some of its functions have side-effects. However, we make use of functions as first-class citizens and closures. By "functions as first-class citizens" we mean that RDF4R has both functions that return functions and functions that take functions as arguments.

The simplest example of such a function is the `query_factory` function which converts a SPARQL query to an R function. We believe that this is a very useful feature because the working ontologist often has quite a few SPARQL queries that they want to execute and then parse the results of. Had we just provided the generic `submit_sparql` function, the workflow for the user would have looked as follows: first, modify the SPARQL query somehow (e.g. by for example changing a label that is matched); second, execute `submit_sparql`—while not forgetting to specify the correct access point—; and, third, parse the results. `query_factory` packages all of this functionality in one place and hides the complexity by allowing the user to simply write

```
genus_lookup("\"Drosophila\"")
```

This example is taken from the previous section. Here, `genus_lookup` is a function that has been dynamically generated by `query_factory` and encloses the functionality for parameterizing the query, executing it, and then parsing the results. This enclosing is possible thanks to the implementation of functions in R as *closures*.

In R functions are implemented as closures: statements of code and a reference to their defining environment[5]. An environment is a data structure that maps names to values. This construction implies that whatever variables were defined in the environment that defined the function are implicitly accessible to the defined function. It is thus possible to encapsulate some of the arguments to the function factory in the constructed function.

We make use of closures in `genus_factory` and even more evidently in the simpler case of `add_data_factory`. `add_data_factory`'s arguments are only the details needed to access a particular endpoint. It returns a function that takes some RDF statements and submits the statements to this endpoint. For example:

```
add_data_to_graphdb = add_data_factory(access_options = graphdb,
    prefixes = prefixes)
add_data_to_graphdb(rdf_data = ttl)
```

The constructed function, `add_data_to_graphdb` does not have the parameter `access_options` any more. Instead, `add_data_to_graphdb` looks for `access_options` in its enclosing environment. This pattern allows us to hide some of the complexity and reduce errors.

Another example of the functional style that we will look at is to be found in the `identifier_factory` and `fidentifier` function. Perhaps, here, a even a further reduction in complexity can be achieved through further efforts. `identifier_factory` takes a list of lookup functions as an input and returns constructor functions. This makes `identifier_factory` into a functor as it both takes functions as inputs and returns functions. The

[5]Please consult Wickham, 2015 for a tutorial on closures and environments.

reasoning behind this functor is to enable the working ontologist to generate code that first looks up a resource identifier in several different places before coining a new one. The syntax achieved as follows:

```
lookup_or_mint_id = identifier_factory(fun = list(simple_lookup),
   prefixes = prefixes,
   def_prefix = eg)

idking_lear = lookup_or_mint_id(list(lking_lear))
```

Here, one has to enclose the arguments to `lookup_or_mind_id` in a list, as it is possible that the SPARQL queries that `lookup_or_mind_id` encapsulates—in this case the single `simple_lookup`—may have more that one parameter. Additional extension of the package are possible in the area of error reporting, as should one forget to enclose `lking_lear` in a list the error message produced is slightly cryptic:

```
> lookup_or_mint_id(lking_lear)
 Show Traceback

 Rerun with Debug
 Error in UseMethod("represent", x) :
  no applicable method for 'represent' applied to an object of class
       "character" In addition: Warning message:
In l$fun = fun : Coercing LHS to a list
```

There are several ways to achieve the desired extension. One is to define a new super class *representable* as a parent class of *literal* and *identifier*). Then extend `lookup_or_mind_id` with check on whether its input is a *representable*.

More in-line with the traditional functional programming style, is to have `lookup_or_mint_id` have a dynamic function signature taking one or more arguments of the *representable* class. I.e. the number of arguments to the function is not fixed. We will provide this extensions in a future release (2.0) of RDF4R.

4.4.3 Elements of Object-Oriented Programming (OOP)

We already briefly touched on the need to define specialized classes in the previous section. Classes may be needed for type-checking, for bundling related functionality together, and for achieving mutable state. There are several ways to implement object-oriented programming in R.

S3

Several functions of RDF4R return lists with their *class* attribute set. The most notable of those are `literal` and `identifier`. There are also several generic functions used to invoke class-specific implementations via a call to `UseMethods`. The most notable of those is `represent`:

```
represent = function(x)
{
     UseMethod("represent", x)
}
```

One uses `represent` by calling it with an either *literal* or *identifier* object (see previous section on Usage). The goal is to make possible the writing of generic code that works with both of literals and of resource identifiers. Whereas, during serialization, literals need to be potentially quoted together with an XSD type (e.g. `"CNN"\textasciicircum\textasciicircum xsd:string`, resource identifiers just need to be pasted in Turtle as they are (e.g. `<http://cnn.com>`). By having this generic function the serialization function does not need to be aware of such details.

R6

We use the R6 classes (Chang, 2017) both for bundling behavior with data for achieving mutable state.

R6 is used for the in-memory representation of RDF (`ResourceDescriptionFormat`). The design decision to use R6 was taken in order to allow users of the package to create their RDF object incrementally, by adding more triples. E.g.

```
classics_rdf = ResourceDescriptionFramework$new()
classics_rdf$add_triple(subject = idshakespeare, predicate = wrote,
    object = idking_lear)
classics_rdf$add_triple(subject = idking_lear, predicate = rdfs_
    label, object = lking_lear)
```

Resizing of R lists is a costly operation if the lists had not been preallocated. Therefore, we have implemented a dynamically reallocating vector `DynVector` as part of the package. `DynVector` initializes a list and every time its length is exceeded by adding new elements, it reallocates itself to double its size. This reallocation results in a constant computation time for the operation addition on average (Harrington, 2018). As we pointed our earlier, however, a future release of RDF4R will support both `DynVector` and `librdf` as in-memory storage models.

Furthermore, we support the `$add_triples(rdf)` method that lets the user grow one `ResourceDescriptionFramework` object with another. Since this method is growing the RDF object by more than one, it can be used in conjunction with a function that returns an RDF object (in the following example `extract_triples`). We can chain the two functions together to obtain all RDF statements that are obtained by application of `extract_triples` to some list:

```
merged_rdf = ResourceDescriptionFramework$new()
lapply(lapply(information, extract_triples), merged_rdf$add_triples)
```

Note that this still can only be executed sequentially in order not to corrupt the in-memory representation of `merged_rdf` as each call to `merged_rdf$add_triples` changes the state of `merged_rdf`. A future release of the package will contain an additional Triple class allowing users to store RDF as lists of triples (and thus benefiting from parallelism constructs such as `parLapply`). The user will have the option of waiting until the last possible moment to create a `ResourceDescriptionFramework` class from a list of triples before serialization.

Chapter 5

Workflows for biodiversity data

In this chapter we discuss two automated workflows for exchange of biodiversity data developed as part of OpenBiodiv: (1) automatic import of specimen records into manuscripts, and (2) automatic generation of data paper manuscripts from Ecological Metadata Language (EML[1]) metadata. The workflows were presented at a webinar[2] for the organization iDigBio[3] and published as a paper (Senderov et al., 2016).

The slides from the presentation as well as a PDF of the paper are available from the webinar GitHub page under https://github.com/vsenderov/idigbio-webinar.

5.1 Abstract

Information on occurrences of species and information on the specimens that are evidence for these occurrences (specimen records) is stored in different biodiversity databases. These databases expose the information via public REST API's. We focused on the Global Biodiversity Information Facility (GBIF), Barcode of Life Data Systems (BOLD), iDigBio, and PlutoF, and utilized their API's to import occurrence or specimen records directly into a manuscript edited in the ARPHA Writing Tool (AWT).

Furthermore, major ecological and biological databases around the world provide information about their datasets in the form of EML. A workflow was developed for creating data paper manuscripts in AWT from EML files. Such files could be downloaded, for example, from GBIF, DataONE, or the Long-Term Ecological Research Network (LTER Network).

5.2 Introduction

We present two workflows developed as part of OpenBiodiv: (1) automatic import of occurrence or specimen records into manuscripts and (2) automatic generation of data paper manuscripts from Ecological Metadata Language (EML) metadata. The aim of the presentation is to familiarize the biodiversity community with these workflows and motivate the workflows from a scientific standpoint.

The development of these workflows focuses on two areas: optimizing the workflow of specimen data and optimizing the workflow of dataset metadata. These efforts resulted in the functionality that it is now possible, via a record identifier, to directly

[1] http://www.dcc.ac.uk/resources/metadata-standards/eml-ecological-metadata-language

[2] Date: 16 June 2017.

[3] Integrated Digitized Biocollections (iDigBio) is a US-based aggregator of biocollections data. They hold regular webinars and workshops aimed at improving biodiversity informatics knowledge, which are attended by collection managers, scientists, and IT personnel. Thus, doing a presentation for iDigBio was an excellent way of making the research and tools-development efforts of OpenBiodiv widely known and getting feedback from the community.

import specimen record information from the Global Biodiversity Information Facility (GBIF), Barcode of Life Data Systems (BOLD), iDigBio, or PlutoF into manuscripts in the ARPHA Writing Tool (AWT). No manual copying or retyping is required. Moreover, we created a second, data paper-based workflow.

The concept of data papers as an important means for data mobilization was introduced to biodiversity science by Chavan and Penev, 2011. The data paper is a scholarly journal publication whose primary purpose is to describe a dataset or a group of datasets, rather than report a research investigation. Data papers serve to increase visibility, provide peer review, permanent scientific record, and credit and citation capabilities (via DOI) for biodiversity data. Thus, data papers support the effort for data to become a first class research product, and are a step forward in the direction of open science (Chavan and Penev, 2011; Chavan, 2013).

Using this workflow, it is now possible to generate a data paper manuscript in AWT from a file formatted in recent EML versions.

5.3 Methods

Both workflows discussed rely on three key standards: RESTful API's for the web (Kurtz, 2013), Darwin Core (Wieczorek et al., 2012), and EML (Fegraus et al., 2005).

RESTful is a software architecture style for the Web, derived from the dissertation of Fielding, 2000. It is out of the scope of this paper to fully explain what a RESTful API is, but a short summary follows (after Kurtz, 2013):

- URI's have to be provided for different resources.
- HTTP verbs have to be used for different actions.
- HATEOAS (Hypermedia as the Engine of Application State) must be implemented. This is a way of saying that the client only needs to have a basic knowledge of hypermedia, in order to use the service.

On the other hand, Darwin Core (DwC) is a standard developed by the Biodiversity Information Standards (TDWG), also known as the Taxonomic Databases Working Group, to facilitate storage and exchange of biodiversity and biodiversity-related information. ARPHA and BDJ use the DwC terms to store taxonomic material citation data.

Finally, EML is an XML-based open-source metadata format developed by the community and the National Center for Ecological Analysis and Synthesis (NCEAS) and the Long Term Ecological Research Network (LTER, Fegraus et al., 2005).

5.3.1 Development of workflow 1: Automated specimen record import

There is some confusion about the terms occurrence record, specimen record, and material citation. A DwC Occurrence[4] is defined as *"an existence of an Organism*[5] *at a particular place at a particular time."* The term specimen record is a term that we use for cataloged specimens in a collection that are evidence for the occurrence. In DwC, the notion of a specimen is covered by MaterialSample[6], LivingSpecimen[7],

[4] http://rs.tdwg.org/dwc/terms/#Occurrence
[5] http://rs.tdwg.org/dwc/terms/#Organism
[6] http://rs.tdwg.org/dwc/terms/#MaterialSample
[7] http://rs.tdwg.org/dwc/terms/#LivingSpecimen

PreservedSpecimen[8], and FossilSpecimen[9]. The description of MaterialSample reads: *"a physical result of a sampling (or sub-sampling) event. In biological collections, the material sample is typically collected, and either preserved or destructively processed."* While there is a semantic difference between an occurrence record (DwC Occurrence) and a specimen record (DwC MaterialSample, LivingSpecimen, PreservedSpecimen, or FossilSpecimen), from the view point of pure syntax, they can be considered equivalent since both types of objects[10] are described by the same fields in our system grouped in the following major groups:

- Record-level information
- Event
- Identification
- Location
- Taxon
- Occurrence
- Geological context

Taxonomic practice dictates that authors cite the materials their analysis is based on in the treatment section of the taxonomic paper (Catapano, 2010). Therefore, in our system, as it is a document-authoring system, we take the view that these objects are material citations, i.e. bibliographic records that refer to a particular observation in the wild or a specimen in a museum. As a Supplementary material 1 to Senderov et al., 2016, we have attached a spreadsheet listing all DwC terms describing a specimen or occurrence record that can be imported into AWT as a material citation. From here on, we will refer to the objects being imported as specimen records, and to the objects that are part of the manuscript as material citations.

At the time when development of the workflow started, AWT already allowed import of specimen records as material citations via manual interface and via spreadsheet (Suppl. material 1 of Senderov et al., 2016). So, the starting point for the development of the workflow was to locate API's for downloading biodiversity and biodiversity-related data from major biodiversity data providers and to transform the data that was provided by these API's into DwC-compatible data, which was then to be imported into the manuscript. We chose to work with GBIF, BOLD Systems, iDigBio and the PlutoF platform.

In Suppl. material 2 of Senderov et al., 2016 we give all the necessary information about the API's and how to map their results to DwC: endpoints, documentation, and the mapping of terms. GBIF and iDigBio name their fields in accordance with DwC, whereas for PlutoF and BOLD Systems there is a direct mapping given in the spreadsheet.

In order to abstract and reuse source code we have created a general Occurrence class, which contains the code that is shared between all occurrences, and children classes GbifOccurrence, BoldOccurrence, IDigBioOccurrence, and PlutoFOccurrence, which contain the provider-specific code. The source code is written in PHP.

[8] http://rs.tdwg.org/dwc/terms/#PreservedSpecimen
[9] http://rs.tdwg.org/dwc/terms/#FossilSpecimen
[10] We are using the term objects here in the computer science sense of the word to denote generalized data structures.

5.3.2 Development of workflow 2: Automated data paper generation

Data papers are scholarly articles describing a dataset or a data package (Chavan and Penev, 2011). Similarly, metadata are data about a dataset (Michener, 2006). Ecological metadata can be expressed in an XML-language called EML (Fegraus et al., 2005). It formalizes the set of concepts needed for describing ecological data (Fegraus et al., 2005). It is broad enough and allows dataset authors from neighboring disciplines (e.g. taxonomy) to annotate their datasets with it. We asked ourselves the question: would it be possible to convert such metadata into a data paper manuscript? As the data paper manuscript in AWT is also stored as XML (for format details see the Pensoft API[11]), all that was needed was an XSLT transformation mapping the fields of EML to the fields in the data paper XML. We have created two such transformations, for EML version 2.1.1 and for EML version 2.1.0, which we've attached as Suppl. material 3 to Senderov et al., 2016. The workflow has been tested with EML metadata downloaded from GBIF, DataONE and the LTER Network, however, it can be used for EML metadata from any other data provider.

5.4 Results and Discussion

5.4.1 Workflow 1: Automated specimen record import into manuscripts developed in the ARPHA Writing Tool

Implementation

It is now possible to directly import a specimen record as a material citation in an ARPHA Taxonomic Paper from GBIF, BOLD, iDigBio, and PlutoF (Slide 5, as well as Fig. 5.1). The workflow from the user's perspective has been thoroughly described in Senderov et al., 2016 and became a part of the routine data publishing workflow of Pensoft described in Penev et al., 2016. In a nutshell, the process works as follows:

1. At one of the supported data portals (BOLD, GBIF, iDigBio, PlutoF), the author locates the specimen record he/she wants to import into the Materials section of a Taxon treatment.

2. Depending on the portal, the user finds either the occurrence identfier of the specimen, or a database record identifier of the specimen record, and copies that into the respective upload field of the ARPHA system (Fig. 5.2).

3. The new material citations are rendered in both human- and machine-readable DwC format in the Materials section of the respective Taxon treatment and can be further edited in AWT, or downloaded from there as a CSV file.

Discussion. The persistent unique identifiers (PID's) are a long-discussed problem in biodiversity informatics (Guralnick et al., 2014). Questions of fundamental importance are: given a specimen in a museum, is it (and how often is it) cited in a paper? What about the quality of the database record belonging to this specimen? In order for us to be able to answer these questions to our satisfaction, specimens need to have their own unique identifiers, which are imported as part of the material citation and allow the researcher to scan through the body of published literature to find which specimens have been cited (and how often). In practice, however, this is not always

[11] Pensoft API available under http://arpha.pensoft.net/dev/.

FIGURE 5.1: This fictionalized workflow presents the flow of information content of biodiversity specimens or biodiversity occurrences from the data portals GBIF, BOLD Systems, iDigBio, and PlutoF, through user-interface elements in AWT to textualized content in a Taxonomic Paper manuscript template intended for publication in the Biodiversity Data Journal.

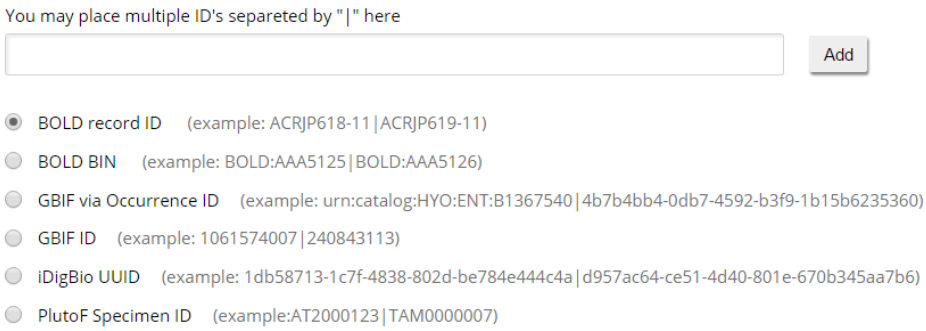


FIGURE 5.2: User interface of the ARPHA Writing Tool controlling the import of specimen records from external databases.

the case and we have to rely on the DwC triplet, (`institutionCode`, `collectionCode`, `catalogNumber`), to positively identify specimens (Guralnick et al., 2014). In the next paragraphs, we discuss how the information provided by the repositories can be used to track material citations.

GBIF. Import from GBIF is possible both via a DwC `occurrenceID`, which is the unique identifier for the specimen/ occurrence, or via a GBIF ID, which is the record ID in GBIF's database. Thanks to its full compliance with DwC, it should be possible to track specimens imported from GBIF.

BOLD Systems. In the BOLD database, specimen records are assigned an identifier, which can look like `ACRJP619-11`. This identifier is the database identifier and is used for the import; it is not the identifier issued to the specimen stored in a given collection. However, some collection identifiers are returned by the API call and are stored in the material citation, for example, DwC catalogNumber and DwC institutionCode (see mappings in Suppl. material 2 in Senderov et al., 2016). In this case, we have what is called a DwC doublet (Guralnick et al., 2014), which provides us with the minimum amount of information to attempt an identification.

A feature of BOLD Systems is that records are grouped into BIN's representing Operational Taxonomic Units (OTU's) based on a hierarchical/ graph-based clustering algorithm (Ratnasingham and Hebert, 2013). It is possible to import all specimen records from a BIN in a single step, specifying the BIN ID. This streamlines the description of new species from OTU's as it allows the taxonomist to save time and import all materials from the BIN.

iDigBio. iDigBio provides its specimen records in a DwC-compatible format. Similar to GBIF, both a DwC occurrenceID, as well as DwC triplet information is returned by the system and stored in our XML making tracking of specimen citations easy.

PlutoF. Import from PlutoF is attained through the usage of a specimen ID (DwC catalogNumber), which is disambiguated to a PlutoF record ID by our system. If a specimen ID matches more than one record in the PlutoF system, multiple records are imported and the user has to delete the superfluous material citations. PlutoF does store a full DwC triplet while no DwC occurrenceID is available for the time being.

Ultimately, this workflow can serve as a curation filter for increasing the quality of specimen data via the scientific peer review process. By importing a specimen record

FIGURE 5.3: Download of an EML from the GBIF Integrated Publishing Toolkit (IPT).

via our workflow, the author of the paper vouches for the quality of the particular specimen record that he or she presumably has already checked against the physical specimen. Then a specimen that has been cited in an article can be marked with a star as a peer-reviewed specimen by the collection manager. Also, the completeness and correctness of the specimen record itself can be improved by comparing the material citation with the database record and synchronizing differing fields.

There is only one component currently missing from for this curation workflow: a query page that accepts a DwC occurrenceID or a DwC doublet/ triplet and returns all the information stored in the Pensoft database regarding material citations of this specimen.

5.4.2 Workflow 2: Automated data paper manuscript generation from EML metadata in the ARPHA Writing Tool

Implementation. We have created a workflow that allows authors to automatically create data paper manuscripts from the metadata stored in EML. The completeness of the manuscript created in such a way depends on the quality of the metadata; however, after generating such a manuscript, the authors can update, edit, and revise it as any other scientific manuscript in the AWT. The workflow has been thoroughly described in a blog post; concise step-wise instructions are available via the publication Penev et al., 2017a In a nutshell, the process is illustrated in Figs. **??** and includes:

1. Obtain metadata formulated as an EML file.
2. Import the metadata into the manuscript via the automated tools.
3. Finalize the manuscript by editing the template created on the basis of the EML file.

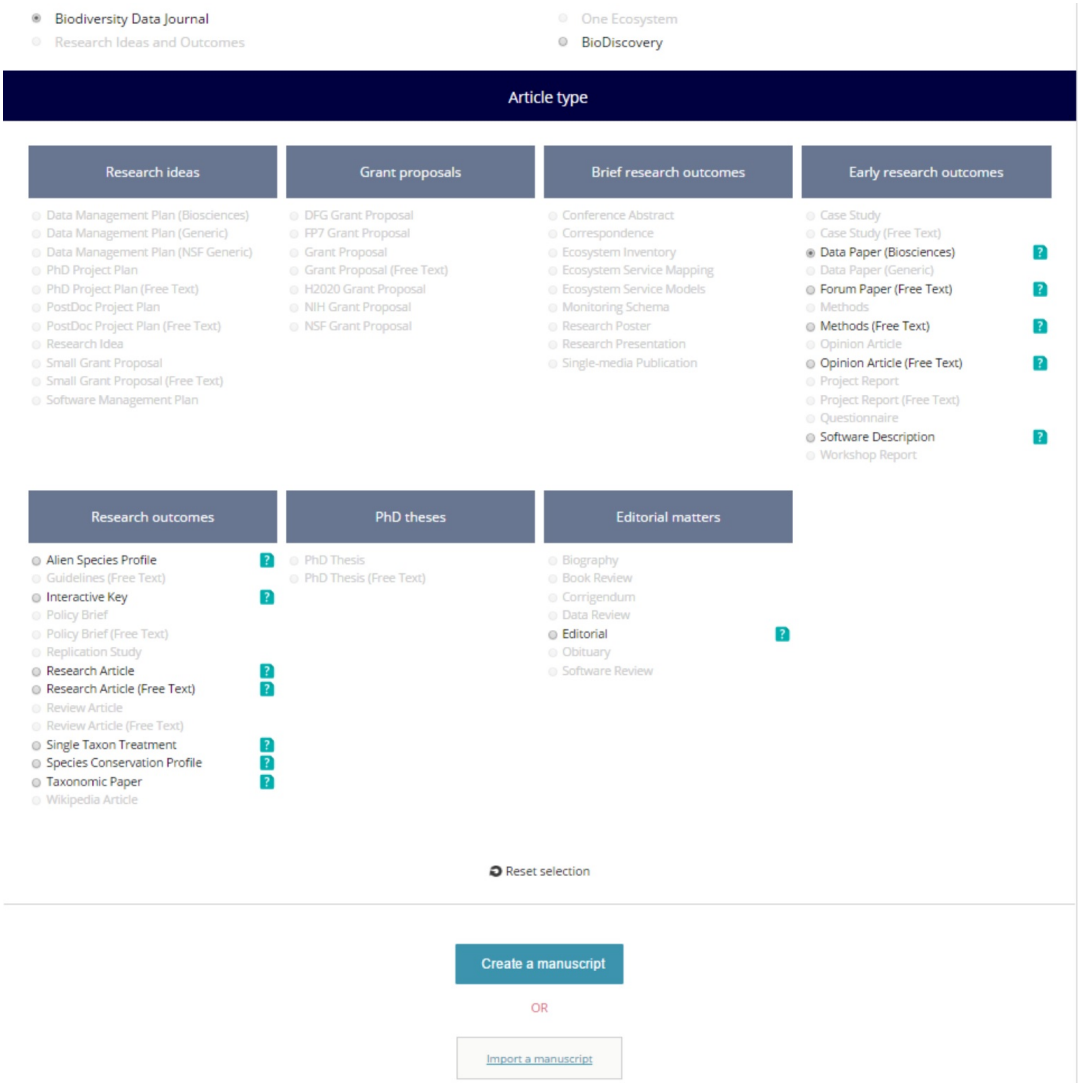


FIGURE 5.4: Selection of the journal and "Data Paper (Biosciences)" template in the ARPHA Writing Tool.

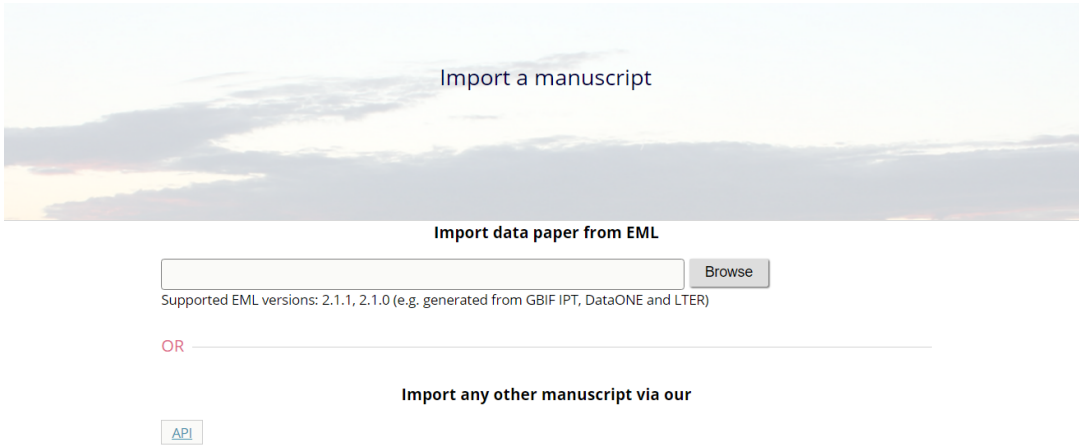


FIGURE 5.5: The user interface field for uploading EML files into ARPHA.

Discussion In 2010, GBIF and Pensoft began investigating mainstream biodiversity data publishing in the form of "data papers." As a result this partnership pioneered a workflow between GBIF's IPT and Pensoft's journals, viz.: ZooKeys, MycoKeys, Phytokeys, Nature Conservation, and others. The rationale behind the project was to motivate authors to create proper metadata for their datasets to enable themselves and their peers to properly make use of the data. Our workflow gives authors the opportunity to convert their extended metadata descriptions into data paper manuscripts with very little extra effort. The workflow generates data paper manuscripts from the metadata descriptions in IPT automatically. Manuscripts are created in Rich Text Format (RTF) format, edited and updated by the authors, and then submitted to a journal to undergo peer review and publication. The publication itself bears the citation details of the described dataset with its own DOI or other unique identifier. Ideally, after the data paper is published and a DOI is issued for it, it should be included in the data description at the repository where the data is stored. Within less than four years, a total of more than 100 data papers have been published in Pensoft's journals (examples: Brosens et al., 2013, Desmet and Brouilet, 2013, Gutt et al., 2013, Pierrat et al., 2012, Shao et al., 2012, Tng et al., 2016). The workflow and associated author guidelines are described in Penev et al., 2016.

The present chapter describes the next technological step in the generation of data papers: direct import of an EML file via an API into a manuscript being written in AWT. A great advantage of the present workflow is that data paper manuscripts can be edited and peer-reviewed collaboratively in the authoring tool even before submission to the journal. These novel features provided by AWT and BDJ may potentially become a huge step forward in experts' engagement and mobilization to publish biodiversity data in a way that facilitates recording, credit, preservation, and re-use. Another benefit of this usage of EML data might be that in the future, more people will provide more robust EML data files.

Feedback. The two workflows presented generated a lively discussion at the end of the presentation, which we summarize below:

1. Are specimen records imported from GBIF and then slightly changed during the editorial process then deduplicated at GBIF? Answer: Unfortunately, no. At GBIF, deduplication only occurs for identical records.

2. Are we leaving the identifiers from GBIF or iDigBio in the records? Answer: Yes. We have made the best effort to import specimen record identifiers. This has been discussed in the previous sections.

3. How will the tool reduce the input time for constructing a manuscript? Answer: AWT reduces the time for creating a manuscript in two significant ways. First of all, the workflows avoid retyping of specimen records or metadata. Secondly, another time-saving feature is the elimination of copying errors. Creating of data paper manuscripts from EML saves, as a minimum, the effort of copy-pasting metadata and their arrangement in a manuscript.

4. What are the major hurdles or challenges left in having this become a main-stream tool? How mature is the tool? Answer: We believe that the main hurdles in this becoming a main-stream tool are visibility and awareness of the tool by the community, as the stability of the software is already at a very good stage.

5. Is it possible to track the usage of museum specimens for data aggregators? Answer: Yes, see question 2 and discussion in the present section.

6. How do you go to the article page where collection managers can search for data published from their collections on the Pensoft website? Answer: We are working on the streamlining of this functionality. It will be part of the OBKMS. Currently, we markup collection codes against the Global Registry of Biodiversity Repositories (GRBio[12]) vocabularies, and the reader can view the records from a particular collection by clicking on the collection code.

[12]GRBio is available under http://grbio.org/

Chapter 6

Web portal

Under openbiodiv.net one can reach the main portal giving access to OpenBiodiv resources. This portal was developed to support OpenBiodiv. OpenBiodiv.net presents two visual elements to the user: the search bar and list of application icons in the bottom. Furthermore, under graph.openbiodiv.net (also accessible from the icon SPARQL endpoint) one can reach the OpenBiodiv workbench, a feature of GraphDB that gives web access to the SPARQL endpoint.

These User Interface (UI) features are designed to facilitate the three user types of the system that we envisage:

1. Basic level: uses search bar.
2. Specialist level: uses apps.
3. Power user: uses the work-bench of the system or R.

6.1 Functionality of the system

6.1.1 Basic usage

The basic level of interaction is for users who want a quick look into the system's database; they can be beginners without knowledge of the Semantic Web or of taxonomy, or advanced users with little time or a very basic query. An example of such a user will simply look for an entity (e.g. taxonomic name, person) and would like to retrieve some information about it. To do so, one simply needs to type the label of the thing that they are looking for. For example, in Fig. 6.1 we typed "Daniel Mietchen" in order to get information about a collaborator of mine. The system automatically identified Daniel as a person (`foaf:Person`) and rendered the RDF statements that it received about him with the appropriate template. The templates are different for the different types (e.g. people, taxa, etc.) and easily customized and extensible. For each resource that is rendered, we always display its stable URI. If one clicks on the URI itself, they will land on the same page.

Note that the system does not need an exact match to display a resource. If we, for example, would misspell Daniel's name and type "Daniel Mitchen" instead, a fuzzy lookup of the Lucene search-index powering the system will find that "Daniel Mietchen" is the closest match to the string we had typed and display the correct page. If there were to be any ambiguity a list of "Alternative interpretation" would be displayed.

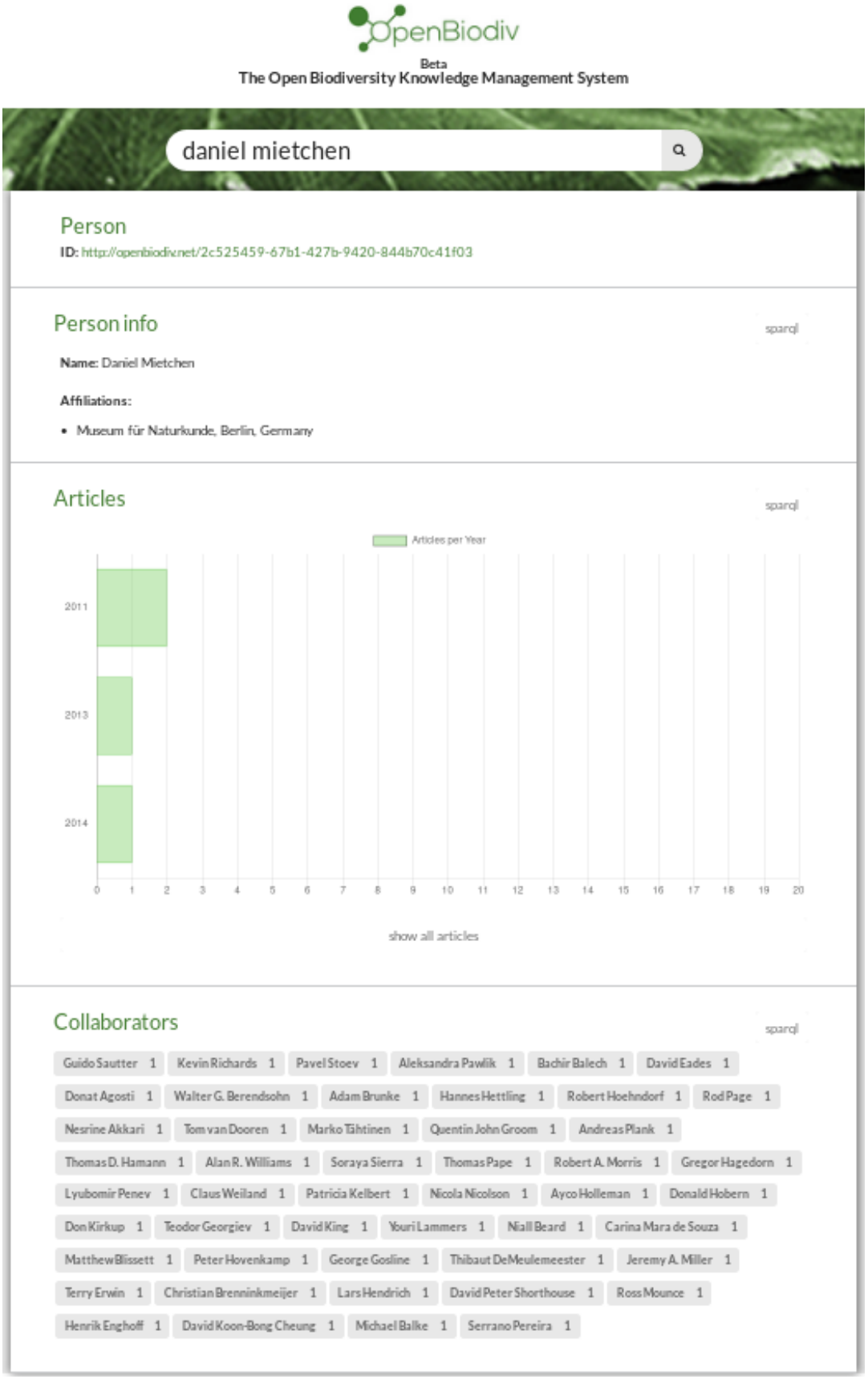


FIGURE 6.1: Illustration of basic usage of OpenBiodiv to look information about a person.

6.1.2 Specialist level

A specialist is someone who has a question of particular taxonomic importance that cannot be answered by a simple name-based look-up. For example, a collection manager at a museum may want to periodically check for articles that make use of their collection in order to justify additional funding to prevent natural disasters. Or a taxonomist interested in a particular region or group may want to stay up to date with published literature fitting those criteria—let's say weevils (Curculionidae) of Arizona, U.S.A. These "filtering" tasks are accomplished in OpenBiodiv's UI-model by semantic apps accessible from the portal. A semantic app is a single-purpose application usually wrapping a UI and visualization component around a SPARQL query. Apps are currently being developed at Pensoft for OpenBiodiv and readers of this manuscript are encouraged to submit their requests to `datascience@pensoft.net`.

6.1.3 Power user

The power user is someone with knowledge of the Semantic Web and its technologies (SPARQL, ontologies, etc.). The power user goes to the workbench and executes their queries there, or uses the functionality of the RDF4R package described in Chapter 4 to execute SPARQL directly on the OpenBiodiv endpoint directly from the R environment.

6.2 Implementation

The UI-components of the web portal are developed in the ReactJS JavaScript framework written by Facebook. Server-side processing is done in PHP. This part of OpenBiodiv is omitted from a detailed discussion in the present dissertation effort.

6.3 Discussion and Outlook

The website is still in beta. The functionality that works great is the semantic search engine. For some basic data types there are templates for visualization. However, the site can not be considered complete and most users use the SPARQL search language.

A future direction, in which the site can be taken, is to expand it with more templates and new apps.

Chapter 7

Listings

This chapter contains source code listings that are too long to be included in-line in the previous chapters.

7.1 Code for the Linked Open Data

Listing 7.1: Taxonomic name usage of the name *P. emarginaticeps* in Taxpub. Name parts are tagged with `tp:taxon-name-part` and the expansion of abbreviations (regularization) is marked up with the attribute `reg`

```
<tp:taxon-name>
  <tp:taxon-name-part taxon-name-part-type="genus" reg="Pristaulacus">
    P.
  </tp:taxon-name-part>
  <tp:taxon-name-part taxon-name-part-type="species" reg="emarginaticeps">
    emarginaticeps
  </tp:taxon-name-part>
  <tp:taxon-name-part taxon-name-part-type="authority">
    Turner 1922
  </tp:taxon-name-part>
</tp:taxon-name>
```

Listing 7.2: Most prolific author SPARQL query.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX openbiodiv: <http://openbiodiv.net/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
select (SAMPLE(?name) AS ?name) (COUNT(DISTINCT ?paper) as ?npapers) where {
        ?author rdf:type foaf:Person ;
            rdfs:label ?name .
    ?paper dcterms:creator ?author .
} GROUP BY ?author
ORDER BY DESC( ?npapers)
```

Listing 7.3: Most mentioned scientific name.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX openbiodiv: <http://openbiodiv.net/>
PREFIX pkm: <http://proton.semanticweb.org/protonkm#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
select (SAMPLE(?name) as ?name) (COUNT(DISTINCT ?tnu) AS ?nmentions) where {
        ?s rdf:type openbiodiv:ScientificName ;
       rdfs:label ?name .
    ?tnu pkm:mentions ?s .
} GROUP BY ?s ORDER BY DESC(?nmentions)
```

Listing 7.4: Most mentioned species name.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX openbiodiv: <http://openbiodiv.net/>
```

```
PREFIX pkm: <http://proton.semanticweb.org/protonkm#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dwciri: <http://rs.tdwg.org/dwc/iri/>
SELECT (MAX(?name) AS ?name) (COUNT(DISTINCT ?tnu) AS ?nmentions) where {
        ?s rdf:type openbiodiv:ScientificName ;
       rdfs:label ?name ;
       dwciri:taxonRank openbiodiv:Species .
    ?tnu pkm:mentions ?s .
} GROUP BY ?s ORDER BY DESC(?nmentions)
```

Listing 7.5: What are the available taxonomic ranks?

```
PREFIX dwciri: <http://rs.tdwg.org/dwc/iri/>
SELECT DISTINCT ?rank
WHERE {
  ?x dwciri:taxonRank ?rank .
}
```

Listing 7.6: Most mentioned species name by number of articles that mention it.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX openbiodiv: <http://openbiodiv.net/>
PREFIX pkm: <http://proton.semanticweb.org/protonkm#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dwciri: <http://rs.tdwg.org/dwc/iri/>
select (MAX(?name) AS ?name) (COUNT(DISTINCT ?a) AS ?narticles) where {
        ?s rdf:type openbiodiv:ScientificName ;
       rdfs:label ?name ;
       dwciri:taxonRank openbiodiv:Species .
    ?tnu pkm:mentions ?s .
    ?a po:contains ?tnu .
    ?a rdf:type fabio:JournalArticle .
} GROUP BY ?s ORDER BY DESC(?narticles)
```

Listing 7.7: Most mentioned scientific name in figures

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX openbiodiv: <http://openbiodiv.net/>
PREFIX pkm: <http://proton.semanticweb.org/protonkm#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dwciri: <http://rs.tdwg.org/dwc/iri/>
PREFIX po: <http://www.essepuntato.it/2008/12/pattern#>
PREFIX fabio: <http://purl.org/spar/fabio/>
PREFIX doco: <http://purl.org/spar/doco/>
select (MAX(?name) AS ?name) (COUNT(DISTINCT ?a) AS ?nmentions) where {
        ?s rdf:type openbiodiv:ScientificName ;
       rdfs:label ?name .
    ?tnu pkm:mentions ?s .
    ?a po:contains ?tnu .
    ?a rdf:type doco:Figure .
} GROUP BY ?s ORDER BY DESC(?nmentions)
```

Listing 7.8: Figures of a given article.

```
# Fetch all the figures (and their captions) belonging
# to an article with a given DOI.

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX openbiodiv: <http://openbiodiv.net/>
PREFIX pkm: <http://proton.semanticweb.org/protonkm#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dwciri: <http://rs.tdwg.org/dwc/iri/>
PREFIX po: <http://www.essepuntato.it/2008/12/pattern#>
PREFIX fabio: <http://purl.org/spar/fabio/>
PREFIX doco: <http://purl.org/spar/doco/>
PREFIX prism: <http://prismstandard.org/namespaces/basic/2.0/>
PREFIX c4o: <http://purl.org/spar/c4o/>
select (GROUP_CONCAT(?caption) AS ?captions) where {
        ?a a fabio:JournalArticle ;
       prism:doi "10.3897/mycokeys.1.1966" .
```

```
    ?f a doco:Figure ;
       c4o:hasContent ?caption .
    ?a po:contains ?f .
} GROUP BY ?a
```

Listing 7.9: Taxonomic discoveries in the weevils.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX openbiodiv: <http://openbiodiv.net/>
PREFIX dwciri: <http://rs.tdwg.org/dwc/iri/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX dwc: <http://rs.tdwg.org/dwc/terms/>
PREFIX pkm: <http://proton.semanticweb.org/protonkm#>
PREFIX prism: <http://prismstandard.org/namespaces/basic/2.0/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
select * where {
    graph openbiodiv:35af6a8a-9817-449e-86dc-dddc81bce09c {
        ?n rdfs:label "Curculionidae" .
        ?c dwciri:scientificName ?n .
        ?s skos:broader ?c .
        ?s dwciri:scientificName ?sn .
    }
     ?sn dwciri:taxonRank openbiodiv:Genus ;
         dwc:genus ?vgenus .
    ?tnu pkm:mentions ?name;
         prism:publicationDate ?date ;
         dwciri:taxonomicStatus openbiodiv:TaxonomicDiscovery .
    ?name dwc:genus ?vgenus;
          rdfs:label ?verbatim .
}
```

Listing 7.10: Sample Lucene query via SPARQL. We have intentionally misspelled the person's name.

```
PREFIX lucene: <http://www.ontotext.com/connectors/lucene#>
PREFIX inst: <http://www.ontotext.com/connectors/lucene/instance#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT *
WHERE {
?search a inst:WordSearch ;
        lucene:query "label:Lubomir␣Penev" ;
        lucene:entities ?resource .

?resource lucene:score ?score ;
   rdfs:label ?label .
} ORDER BY DESC (?score)
```

Listing 7.11: Asks if the name given by the label has been replaced.

```
PREFIX : <http://openbiodiv.net/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
ASK {
  ?name rdf:type :ScientificName ;
        rdfs:label "Pompilidae" .

  ?name :replacementName ?replacementName .

  FILTER NOT EXISTS {?replacementName :replacementName ?anotherName .}
}
```

Listing 7.12: Asks if the name given by the label is considered unavailable.

```
PREFIX : <http://openbiodiv.net/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX pkm: <http://proton.semanticweb.org/protonkm#>
```

```
PREFIX dwciri: <http://rs.tdwg.org/dwc/iri/>
PREFIX openbiodiv: <http://openbiodiv.net/>
PREFIX prism: <http://prismstandard.org/namespaces/basic/2.0/>
PREFIX po: <http://www.essepuntato.it/2008/12/pattern#>
PREFIX fabio: <http://purl.org/spar/fabio/>
ASK
{
  ?tnu pkm:mentions ?name .
  ?name rdfs:label "Messerschmidia␣incana␣G.␣Mey.␣1818" .
  ?tnu dwciri:taxonomicStatus openbiodiv:UnavailableName .
  ?tnu prism:publicationDate ?date .

  # potential revalidation
  FILTER NOT EXISTS {
    ?tnu2 pkm:mentions ?name .
    ?tnu2 dwciri:taxonomicStatus openbiodiv:AvailableName .
    ?tnu2  prism:publicationDate ?date2 .
    FILTER (?date2 > ?date)
  }
}
```

Listing 7.13: Impact of fire in Museu Nacional on biodiversity knowledge.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX openbiodiv: <http://openbiodiv.net/>
PREFIX pkm: <http://proton.semanticweb.org/protonkm#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX po: <http://www.essepuntato.it/2008/12/pattern#>
PREFIX : <http://www.ontotext.com/connectors/lucene#>
PREFIX fabio: <http://purl.org/spar/fabio/>
PREFIX dcelements: <http://purl.org/dc/elements/1.1/>
PREFIX dwc: <http://rs.tdwg.org/dwc/terms/>
PREFIX prism: <http://prismstandard.org/namespaces/basic/2.0/>
PREFIX dwciri: <http://rs.tdwg.org/dwc/iri/>

SELECT ?institution_name (COUNT( DISTINCT ?icu) AS ?times_mentioned)
(COUNT(DISTINCT ?title) AS ?articles)
(GROUP_CONCAT(DISTINCT ?title; SEPARATOR=",␣") AS ?doi_of_articles)
(GROUP_CONCAT(DISTINCT ?name; SEPARATOR=",␣") AS ?names_mentioned)
(COUNT (DISTINCT ?name) AS ?number_of_taxa)
{
        ?s rdf:type openbiodiv:ScientificName ;
       rdfs:label ?name .
    ?tnu pkm:mentions ?s .

    BIND("Museu␣Nacional␣de␣Rio␣de␣Janeiro␣(MNRJ)" as ?institution_name)
        ?icu rdf:type openbiodiv:InstitutionalCodeUsage ;
         dwc:institutional_code "MNRJ" .
    ?container po:contains ?icu, ?tnu ;
               rdf:type fabio:JournalArticle ;
                        prism:doi ?title .
} GROUP BY ?institution_name
ORDER BY DESC (?times_mentioned)
```

Listing 7.14: XML snippet of an author.

```
<contrib contrib-type="author" corresp="no">
  <name name-style="western">
    <surname>Wachkoo</surname>
    <given-names>Aijaz Ahmad</given-names>
  </name>
  <uri content-type="orcid">https://orcid.org/0000-0003-2506-9840</uri>
  <xref ref-type="aff" rid="A3">3</xref>
</contrib>

<aff id="A3">
 <label>3</label>
 <addr-line>
   Central Institute of Temperate Horticulture, Srinagar, Jammu & Kashmir, India
 </addr-line>
</aff>
```

Listing 7.15: RDF snippet of an author. This is a somewhat idealized situation in which the language of the address was available from the article.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

:a a foaf:Person ;
   rdfs:label "Aijaz␣Ahmad␣Wachkoo".
   :affiliation "Central␣Institute␣of␣Temperate␣Horticulture,␣Srinagar,␣Jammu␣&␣Kashmir,␣India"@en ;
   foaf:familyName "Wachkoo" ;
   foaf:givenName "Aijaz␣Ahmad" .
```

Listing 7.16: .

```
:2b836ad5-db56-4093-9752-33c9f7892de6   rdf:type   fabio:JournalArticle ;
  rdfs:label   "Changes␣to␣publication␣requirements␣made␣at␣the␣XVIII␣Internation\
al␣Botanical␣Congress␣in␣Melbourne␣-␣what␣does␣e-publication␣mean␣for␣you?" ;
  dc:title   "Changes␣to␣publication␣requirements␣made␣at␣the␣XVIII␣International\
␣Botanical␣Congress␣in␣Melbourne␣-␣what␣does␣e-publication␣mean␣for␣you?" ;
 prism:doi   "10.3897/mycokeys.1.1961" ;
 dc:publisher   "Pensoft␣Publishers" ;
 prism:publicationDate   "2011-9-14"^^xsd:date ;
 dcterms:publisher   openbiodiv:0df76aab-1fcf-4118-8e50-198e830a7bed .
 openbiodiv:151a37ba-a337-4855-8e01-200f5ec0251b   rdf:type   deo:Introduction ;
         po:isContainedBy   openbiodiv:2b836ad5-db56-4093-9752-33c9f7892de6 .
}
```

Listing 7.17: Update rule for replacement name.

```
PREFIX dwciri: <http://rs.tdwg.org/dwc/iri/>
PREFIX openbiodiv: <http://openbiodiv.net/>
PREFIX po: <http://www.essepuntato.it/2008/12/pattern#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX pkm: <http://proton.semanticweb.org/protonkm#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

INSERT
{
    GRAPH <http://openbiodiv.net/Updates>
    {
    ?name2 openbiodiv:replacementName ?name .
   }
}

WHERE {
        ?tnu1 dwciri:taxonomicStatus openbiodiv:ReplacementName ;
       pkm:mentions ?name.
     ?name dwciri:taxonRank ?rank;
           rdfs:label ?vname .

    ?s po:contains ?tnu .
    ?s po:contains ?citations.
    ?citations rdf:type openbiodiv:NomenclatureCitationsList;
               po:contains ?tnu2 .
    ?tnu2 rdf:type openbiodiv:TaxonomicNameUsage ;
          pkm:mentions ?name2.
    ?name2 rdfs:label ?vname2;
           dwciri:taxonRank ?rank.
}
```

Listing 7.18: Update rule for related name.

```
PREFIX dwciri: <http://rs.tdwg.org/dwc/iri/>
PREFIX : <http://openbiodiv.net/>
PREFIX po: <http://www.essepuntato.it/2008/12/pattern#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX pkm: <http://proton.semanticweb.org/protonkm#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

INSERT
{
```

```
    GRAPH <http://openbiodiv.net/Updates>
    {
    ?name2 :relatedName ?name .
   }
}

WHERE {
  ?nom_sec rdf:type :NomenclatureSection ;
    :contains ?tnu1 .

  ?tnu1 rdf:type :TaxonomicNameUsage ;
    pkm:mentions ?name.

  ?nom_sec :contains ?tnu2 .

  ?tnu2 rdf:type :TaxonomicNameUsage ;
    pkm:mentions ?name2.

  FILTER(?name != ?name2)
}
```

7.2 Code for the R Library

Listing 7.19: R code: connecting to an RDF database using RDF4R. Outputs are given as comments after the statements.

```
library(rdf4r)

openbiodiv = rdf4r::basic_triplestore_access(
  server_url = "http://graph.openbiodiv.net",
  repository = "depl2018-lite"
)

graphdb = rdf4r::basic_triplestore_access(
  server_url = "http://graph.openbiodiv.net",
  user = "dbuser",
  password = "public-access",
  repository = "test"
)

graphdb
# $server_url
# [1] "http://graph.openbiodiv.net"
# $repository
# [1] "test"
# $authentication
# <request>
# Options:
# * httpauth: 1
# * userpwd: dbuser:public-access
# $status
# [1] 8
# attr(,"class")
# [1] "list"                          "triplestore_access_options"

openbiodiv
# $server_url
# [1] "http://graph.openbiodiv.net"
# $repository
# [1] "depl2018-lite"
# $authentication
# NULL
# $status
# [1] 8
# attr(,"class")
# [1] "list"                          "triplestore_access_options"

get_protocol_version(graphdb)
# [1] 8

list_repositories(graphdb)
```

```
# uri             id                                                    readable writable
# 1 http://graph.openbiodiv.net/repositories/SYSTEM         SYSTEM    true     true
# 2 http://graph.openbiodiv.net/repositories/depl2018-mini2 depl2018-mini2     true
true
# 3 http://graph.openbiodiv.net/repositories/depl2018       depl2018  true     true
# 4 http://graph.openbiodiv.net/repositories/test           test      true     true
# 5 http://graph.openbiodiv.net/repositories/depl2018-lite  depl2018-lite      true
true
```

Listing 7.20: R. Parameterized SPARQL query to lookup a genus in OpenBiodiv.

```
PREFIX dwciri: <http://rs.tdwg.org/dwc/iri/>
PREFIX openbiodiv: <http://openbiodiv.net/>
PREFIX po: <http://www.essepuntato.it/2008/12/pattern#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX pkm: <http://proton.semanticweb.org/protonkm#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

INSERT
{
    GRAPH <http://openbiodiv.net/Updates>
    {
    ?name2 openbiodiv:replacementName ?name .
   }
}

WHERE {
        ?tnu1 dwciri:taxonomicStatus openbiodiv:ReplacementName ;
       pkm:mentions ?name.
     ?name dwciri:taxonRank ?rank;
           rdfs:label ?vname .

    ?s po:contains ?tnu .
    ?s po:contains ?citations.
    ?citations rdf:type openbiodiv:NomenclatureCitationsList;
               po:contains ?tnu2 .
    ?tnu2 rdf:type openbiodiv:TaxonomicNameUsage ;
          pkm:mentions ?name2.
    ?name2 rdfs:label ?vname2;
           dwciri:taxonRank ?rank.
}

genus_lookup("\"Drosophila\"")
#    genus title
# 1   Drosophila  Characterisation of the chemical profiles of Brazilian and Andean
# morphotypes belonging to the Anastrephafraterculus complex (Diptera, Tephritidae)
# 2   Drosophila                A new species group in the genus Dichaetophora, with
# descriptions of six new species from the Oriental region (Diptera, Drosophilidae)
```

Listing 7.21: R. Literal construction.

```
lking_lear      = literal(text_value = "King␣Lear",        lang = "en")
las_you_like_it = literal(text_value = "As␣You␣Like␣It",   lang = "en")
lhamlet         = literal(text_value = "Hamlet",           lang = "en")
lothello        = literal(text_value = "Othello",          lang = "en")
lsonnet_78      = literal(text_value = "Sonnet␣78",        lang = "en")
lastrophil      = literal(text_value = "Astrophil␣and␣Stella",
lang = "en")
ledward2        = literal(text_value = "Edward␣II",        lang = "en")
lhero           = literal(text_value = "Hero␣and␣Leander", lang = "en")
lgreensleeves   = literal(text_value = "Greensleeves",     lang = "en")
lshakespeare         = literal(text_value = "Shakespeare")
lsir_phillip_sidney  = literal(text_value = "Sir␣Phillip␣Sidney")
lchristopher_marlowe = literal(text_value = "Christopher␣Marlowe")
lhenry_8_rex         = literal(text_value = "Henry␣VII␣Rex")
l1599 = literal(text_value = "1599", xsd_type = xsd_integer)
l1603 = literal(text_value = "1603", xsd_type = xsd_integer)
l1609 = literal(text_value = "1609", xsd_type = xsd_integer)
l1590 = literal(text_value = "1590", xsd_type = xsd_integer)
l1592 = literal(text_value = "1592", xsd_type = xsd_integer)
```

```
l1593 = literal(text_value = "1593", xsd_type = xsd_integer)
l1525 = literal(text_value = "1525", xsd_type = xsd_integer)
```

Listing 7.22: R. Representation of literals.

```
lhamlet
# [1] "\"Hamlet\"@en"

str(lhamlet)
# List of 4
# $ text_value: chr "Hamlet"
# $ xsd_type  :List of 4
# $ lang      : chr "en"
# $ squote    : chr "\"Hamlet\"@en"
# - attr(*, "class")= chr "literal"

represent(lhamlet)
# [1] "\"Hamlet\"@en"

lshakespeare
# [1] "\"Shakespeare\""

str(lshakespeare)
# List of 4
# $ text_value: chr "Shakespeare"
# $ xsd_type  :List of 4
# $ lang      : chr ""
# $ squote    : chr "\"Shakespeare\""
# - attr(*, "class")= chr "literal"

represent(lshakespeare)
# [1] "\"Shakespeare\""

l1603
# [1] "\"1603\"^^xsd:integer"

str(l1603)
# List of 4
# $ text_value: chr "1603"
# $ xsd_type  :List of 4
# $ lang      : chr ""
# $ squote    : chr "\"1603\"^^xsd:integer"
# - attr(*, "class")= chr "literal"

represent(l1603)
# [1] "\"1603\"^^xsd:integer"
```

Listing 7.23: R. Entering identifiers.

```
prefixes = c(
   rdfs = "http://www.w3.org/2000/01/rdf-schema#",
   rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
   example = "http://rdflib-rdf4r.net/",
   art = "http://art-ontology.net/"
 )
eg = prefixes[3]
art = prefixes[4]
artist = identifier(id = "Artist", prefix = art)
play = identifier(id = "Play", prefix = art)
poem = identifier(id = "Poem", prefix = art)
song = identifier(id = "Song", prefix = art)
wrote = identifier(id = "wrote", prefix = art)
has_year = identifier(id = "has_year", prefix = art)
```

Listing 7.24: R. Entering identifiers.

```
artist
# [1] "art:Artist"

str(artist)
# List of 4
```

```
# $ id     : chr "Artist"
# $ uri    : chr "<http://art-ontology.net/Artist>"
# $ qname  : chr "art:Artist"
# $ prefix: Named chr "http://art-ontology.net/"
#  ..- attr(*, "names")= chr "art"
# - attr(*, "class")= chr "identifier"

represent(artist)
# [1] "<http://art-ontology.net/Artist>"

wrote
# [1] "art:wrote"

str(wrote)
# List of 4
# $ id     : chr "wrote"
# $ uri    : chr "<http://art-ontology.net/wrote>"
# $ qname  : chr "art:wrote"
# $ prefix: Named chr "http://art-ontology.net/"
#  ..- attr(*, "names")= chr "art"
# - attr(*, "class")= chr "identifier"

represent(wrote)
# [1] "<http://art-ontology.net/wrote>"
```

Listing 7.25: R. Identifier factory.

```
p_query = "SELECT␣DISTINCT␣?id␣WHERE␣{
␣␣?id␣rdfs:label␣%label
}"

simple_lookup = query_factory(p_query, access_options = graphdb)

lookup_or_mint_id = identifier_factory(fun = list(simple_lookup),
   prefixes = prefixes,
   def_prefix = eg)

idking_lear = lookup_or_mint_id(list(lking_lear))
idas_you_like_it = lookup_or_mint_id(list(las_you_like_it))
idhamlet = lookup_or_mint_id(list(lhamlet))
idothello = lookup_or_mint_id(list(lothello))
idsonnet78 = lookup_or_mint_id(list(lsonnet_78))
idastrophil = lookup_or_mint_id(list(lastrophil))
idedward2 = lookup_or_mint_id(list(ledward2))
idhero = lookup_or_mint_id(list(lhero))
idgreensleeves = lookup_or_mint_id(list(lgreensleeves))
idshakespeare = lookup_or_mint_id(list(lshakespeare))
idsir_phillip_sidney = lookup_or_mint_id(list(lsir_phillip_sidney))
idchristopher_marlowe = lookup_or_mint_id(list(lchristopher_marlowe))
idlhenry_8_rex = lookup_or_mint_id(list(lhenry_8_rex))
```

Listing 7.26: R. Creating RDF.

```
classics_rdf = ResourceDescriptionFramework$new()

classics_rdf$add_triple(subject = idshakespeare,
    predicate = wrote,      object = idking_lear
)
classics_rdf$add_triple(subject = idking_lear,
    predicate = rdfs_label, object = lking_lear
)
classics_rdf$add_triple(subject = idshakespeare,
    predicate = wrote,      object = idas_you_like_it
)
classics_rdf$add_triple(subject = idas_you_like_it,
    predicate = rdfs_label, object = las_you_like_it
)
classics_rdf$add_triple(subject = idas_you_like_it,
    predicate = has_year,   object = l1599
)
classics_rdf$add_triple(subject = idas_you_like_it,
    predicate = rdf_type,   object = play
```

```
)
```

Listing 7.27: Creating RDF

```
classics_rdf$set_context(identifier(id = "classic_example", prefix  = eg))
cat(classics_rdf$serialize())
# @prefix example: <http://rdflib-rdf4r.net/> .
# @prefix art: <http://art-ontology.net/> .
# @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
# @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
# example:classic_example {
# example:0bac3b32-8aac-11e8-a1c9-c961fe5afe72   art:wrote
#  example:0aeae996-8aac-11e8-a1c9-c961fe5afe72 ,  example:0aff274e-8aac-11e8-a1c9-c961fe5afe72
# example:0aeae996-8aac-11e8-a1c9-c961fe5afe72   rdfs:label   "King Lear"@en .
# example:0aff274e-8aac-11e8-a1c9-c961fe5afe72   rdfs:label   "As You Like It"@en ;
#         art:has_year   "1599"^^xsd:integer ;
#         rdf:type   art:Play .
# }
```

Listing 7.28: Creating RDF

```
# via add_data
add_data(classics_rdf$serialize(), access_options = graphdb)
simple_lookup(represent(lking_lear))
# id
# 1 http://rdflib-rdf4r.net/0aeae996-8aac-11e8-a1c9-c961fe5afe72
simple_lookup(represent(lking_lear))
# id
# 1 http://rdflib-rdf4r.net/0aeae996-8aac-11e8-a1c9-c961fe5afe72
p_query_describe = "PREFIX␣example:␣<http://rdflib-rdf4r.net/>
+␣SELECT␣?p␣?o
+␣WHERE␣{
+␣%resource␣?p␣?o␣.
+␣}"
describe = query_factory(p_query = p_query_describe, access_options = graphdb)
describe(represent(idshakespeare))
# p                                                    o
# 1 http://art-ontology.net/wrote http://rdflib-rdf4r.net/0aeae996-8aac-11e8-a1c9-c961fe5afe72
#2 http://art-ontology.net/wrote http://rdflib-rdf4r.net/0aff274e-8aac-11e8-a1c9-c961fe5afe72
describe(represent(idas_you_like_it))
#                 p                          o
# 1 http://www.w3.org/1999/02/22-rdf-syntax-ns#type http://art-ontology.net/Play
# 2      http://www.w3.org/2000/01/rdf-schema#label               As You Like It
# 3                http://art-ontology.net/has_year                         1599
```

Chapter 8

Addendum

8.1 iDigBio presentation

This section contains details of the webinar accompanying Chapter 5.

A video recording of the presentation is available from iDigBio. More information can be found in the webinar information page over at iDigBio. The slides of the presentation are attached as supplementary files and are deposited in Slideshare[1].

During the presentation we conducted a poll about the occupation of the attendees, the results of which are summarized in Fig. 8.1. Of the participants who voted, about a half were scientists, mostly biologists, while the remainder were distributed across IT specialists and librarians, with 20% "Other." The other categories might have been administrators, decision-makers, non-biology scientists, collections personnel, educators, etc.

At the end of the presentation, very interesting questions were raised and discussed. For details, see the "Results and discussion" section of this paper.

Larry Page, Project Director at iDigBio, wrote: "This workflow has the potential to be a huge step forward in documenting use of collections data and enabling iDigBio and other aggregators to report that information back to the institutions providing the data."

Neil Cobb, a research professor at the Department of Biological Sciences at the Northern Arizona University, suggested that the methods, workflows and tools addressed during the presentation could provide a basis for a virtual student course in biodiversity informatics.

[1]The electronic version of the PDF contains clickable links to the video, notes, and slides.

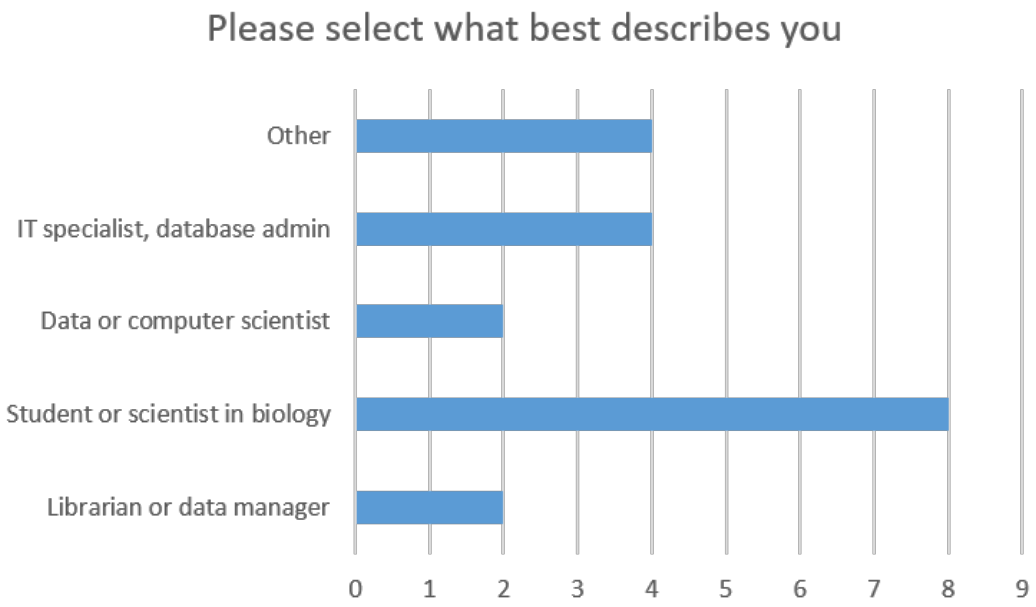


FIGURE 8.1: Poll results about composition of audience during live participation..

Conclusion

Results

We believe that the presented scientific work fulfills the stated objective and tasks.

Result 1. The central result of the thesis is the creation of a domain conceptualization of biodiversity publishing and a formal ontology OpenBiodiv-O enabling the linking of biodiversity knowledge on the basis of scholarly publications. This result has been described in Chapter 2 and in Senderov et al., 2018 and fulfills Objective 1. The source code of the ontology is available under github.com/pensoft/openbiodiv-o.

Result 2. The second result of the thesis is the creation of the software architecture of the OpenBiodiv system outlined in Chapter 1 and Senderov and Penev, 2016. This result fulfills Objective 2.

Result 3. The third result of the thesis has been the creation of a Linked Open Dataset, OpenBiodiv-LOD, consisting of a transformation to RDF-triples and integration in a single store of information from three major repositories of biodiversity data: the XML sources of biological journals published by Pensoft Publishers, the XML sources of treatments freed by Plazi, and a CSV dump of GBIF's taxonomic backbone. OpenBiodiv-LOD is available under `graph.openbiodiv.net` and has been described in Chapter 3. This result fulfills Objective 3.

Result 4. In order to create the Linked Open Data, a software package for the R programming environment, RDF4R, was developed. RDF4R enables the manipulation of RDF data within R and facilities the transformation of scientific publications from a semi-structured XML format to structured semantic RDF. This result has been discussed in Chapter 4 and fulfills Objective 4. The package is available online as free software under `github.com/pensoft/rdf4r`. Furthermore, additional source code (unoptimized) describing XML schemata of Pensoft and Plazi and working in tandem with RDF4R to convert XML to RDF can be found under `github.com/pensoft/ropenbio`.

Result 5. The mechanisms to convert semi-structured XML into RDF-triples are complemented by workflows enabling the enrichment of the XML sources of Pensoft journals by data automatically imported from the major international biodiversity data repositories: BOLD, GBIF, iDigBio, as well as PlutoF. Furthermore, it is now possible, thanks to this dissertation effort to automatically create manuscripts from metadata encoded in the Ecological Metadata Language (EML). The discussion of these automated workflows—automatic data paper generation and automatic occurrence record import—is carried out in Chapter 5. It fulfills Objective 5.

Result 6. To complement the creation of OpenBiodiv-LOD, we have developed a website running on top of the knowledge graph openbiodiv.net, containing a semantic search engine and apps. The website is discussed in Chapter 6 and fulfills Objective 6.

Discussion, conclusion, and outlook

OpenBiodiv-O serves as the basis of the Linked Open Data OpenBiodiv-LOD. By developing an ontology focusing on biological taxonomy, we provided an ontology that fills in the gaps between ontologies for biodiversity resources such as Darwin-SW and semantic publishing ontologies such as the ontologies comprising the SPAR Ontologies. Moreover, we take the view that it is advantageous to model the taxonomic process itself rather than any particular state of knowledge. At this stage, the coverage of the ontology of the different types of resources is sufficient to be the basis for creating the LOD. In this sense, it is completed. On the other hand, adding classes and properties for new types of biodiversity data is possible and desirable.

The LOD, similar to the ontology, are already a solid resource for biologists, as they include information from most articles published by Pensoft and Plazi and count over 600 million triplets. Like the ontology, they should be expanded.

Since the RDF4R package was successfully used to create an LOD, it can be considered complete. Like any software package, however, it should be maintained and developed.

The website is still in beta. The functionality that works great is the semantic search engine. For some basic data types there are templates for visualization. However, the site can not be considered complete and most users use the SPARQL search language.

An important conclusion that can be drawn from the work is that it is possible to use a semantic graph for the integration of a large volume of data on biodiversity. We were unexpectedly given the opportunity to illustrate the power of the knowledge graph by analyzing the damage from the tragic fire at the Museu Nacional in Rio de Janeiro. In addition, we have illustrated that it is possible to write relatively simple logical conclusions to check the validity of a taxonomic name.

Due to the large amount of data, we found that although the use of a semantic graph was possible, some of the initially chosen technologies proved to be inapplicable or difficult to apply. We have observed (see Chapter 3) that the practical application of the full logical OWL model is difficult due to performance problems. Instead in the end, we utilized RDFS that is less powerful but faster. Another observation of ours is that although the R programming environment has given us some advantages in rapidly creating the prototype of the system, by increasing the complexity of the program code needed in the real-life system to cover all private cases, a language with dynamic types such as R creates headaches in debugging. At the same time, we were impressed by the powerful functional programming toolkit R provided.

A big difficulty was the disambiguation of resources such as author names or taxonomic names. In the functional design of the RDF4R package we have put modules that allow us to insert a list of functions/rules for disambiguation when searching for an identifier for a given resource. However, we had only limited success with the rule-based disambiguation and for this reason in the production system it was discontinued at the moment.

Considering these and other "lessons," the future development of the OpenBiodiv project can be outlined in the following not necessarily comprehensive way:

1. As an immediate goal, to expand the LOD and ontology with new data types and new data sources using the existing framework. Such data are e.g. genomic data, occurrence data, (bio-)geographic data, visual data, descriptive data, etc.

2. Look for even closer integration with other existing biodiversity data repositories than GBIF. For example, BioImages, iNaturalist, BOLD, and so on.

3. As a longer-term task to study the transition from a semantic graph to a technology where the inference engine is separated from the data base layer as WikiData or Neo4j. In addition to increased performance, this will give extra flexibility to the project, such as allowing the use of non-RDF-based inference engines such as Euler.

4. Continue developing system software with an even wider application of functional programming and porting it into a functional language like, for example, Haskell or O'CAML.

5. To investigate the problem of disambiguation and related problems for named entity recognition of interesting resources from biodiversity, as well various image recognition tasks, from the point of view of machine learning.

6. Expanding the website with more templates and new applications.

Key scientific and applied contributions

The results discussed in the previous two sections determine the following scientific and applied contributions:

1. Scientific contribution: creating an ontology and a formal model of the field of biodiversity knowledge publication.

2. Applied scientific contribution: analyzing information sources and Creating OpenBiodiv-LOD.

3. Applied scientific Contribution: the implementation of OpenBiodiv software modules.

Our ontology fills the unique niche between bibliographic ontologies such as SPAR and ontologies for biodiversity such as Darwin-SW and as such is undoubtedly of great scientific interest to the biodiversity informatics community. The work has a serious scientific and applied character by providing both a Linked Open Dataset on top of the ontology and software for its users and system developers.

List of publications

Publications in international scientific journals

1. Viktor Senderov and Lyubomir Penev. 2016. "The Open Biodiversity Knowledge Management System in Scholarly Publishing". *Research Ideas and Outcomes* 2, no. e7757 (). ISSN: 2367-7163. doi:10.3897/rio.2.e7757. http://rio.pensoft.net/articles.php?id=7757. Unique citations by Franz and Sterner, 2018, Ordynets et al., 2017 and Burt and Mengual, 2017.

2. Sarah Faulwetter, Evangelos Pafilis, Lucia Fanini, Nicolas Bailly, Donat Agosti, Christos Arvanitidis, Laura Boicenco, Terry Catapano, Simon Claus, Stefanie Dekeyzer, Teodor Georgiev, Aglaia Legaki, Dimitra Mavraki, Anastasis Oulas, Gabriella Papastefanou, Lyubomir Penev, Guido Sautter, Dmitry Schigel, Viktor Senderov, Adrian Teaca, and Marilena Tsompanou. 2016. "EMODnet Workshop on mechanisms and guidelines to mobilise historical data into biogeographic databases". *Research Ideas and Outcomes* 2, no. e10445 (). ISSN: 2367-7163. doi:10.3897/rio.2.e10445. http://rio.pensoft.net/articles.php?id=10445. Unique citation by Pyron, 2018.

3. Pedro Cardoso, Pavel Stoev, Teodor Georgiev, Viktor Senderov, and Lyubomir Penev. 2016. "Species Conservation Profiles compliant with the IUCN Red List of Threatened Species". *Biodiversity Data Journal* 4 (): e10356. ISSN: 1314-2828, 1314-2836. doi:10.3897/BDJ.4.e10356. http://bdj.pensoft.net/articles.php?id=10356. Indexed in WoS SCOPUS, as well as SJR 0.465. Unique citations by Bachman et al., 2018, Lin et al., 2017, Li et al., 2017, Milano et al., 2017.

4. Viktor Senderov, Teodor Georgiev, and Lyubomir Penev. 2016. "Online direct import of specimen records into manuscripts and automatic creation of data papers from biological databases". *Research Ideas and Outcomes* 2 (): e10617. ISSN: 2367-7163. doi:10.3897/rio.2.e10617. http://rio.pensoft.net/articles.php?id=10617. Unique citations by Ordynets et al., 2017.

5. Lyubomir Penev, Daniel Mietchen, Vishwas Chavan, Gregor Hagedorn, Vincent Smith, David Shotton, Éamonn Ó Tuama, Viktor Senderov, Teodor Georgiev, Pavel Stoev, Quentin Groom, David Remsen, and Scott Edmunds. 2017b. "Strategies and guidelines for scholarly publishing of biodiversity data". *Research Ideas and Outcomes* 3, no. e12431 (). ISSN: 2367-7163. doi:10.3897/rio.3.e12431. http://riojournal.com/articles.php?id=12431 Unique citations by Tennant et al., 2017, Marwick and Birch, 2017, Kissling et al., 2018, Mathieu, 2018, Шашков and Иванова, 2018, Шашков et al., 2017, Filippova et al., 2017, Филиппова et al., 2017.

6. Lyubomir Penev, Teodor Georgiev, Peter Geshev, Seyhan Demirov, Viktor Senderov, Iliyana Kuzmova, Iva Kostadinova, Slavena Peneva, and Pavel Stoev. 2017a. "ARPHA-BioDiv: A toolbox for scholarly publication and dissemination of biodiversity data based on the ARPHA Publishing Platform". *Research Ideas and Outcomes* 3, no. e13088 (). ISSN: 2367-7163. doi:10.3897/rio.3.e13088. http://riojournal.com/articles.php?id=13088

7. Emmanuel Arriaga-Varela, Matthias Seidel, Albert Deler-Hernández, Viktor Senderov, and Martin Fikácek. 2017. "A review of the Cercyon Leach (Coleoptera, Hydrophilidae, Sphaeridiinae) of the Greater Antilles". *ZooKeys* 681 (): 39–93. ISSN: 1313-2970, 1313-2989. doi:10.3897/zookeys.681.12522. https://zookeys.pensoft.net/articles.php?id=12522. Indexed in WoS IF 1.079, Q3 SCOPUS, SJR 0.533.

8. Viktor Senderov, Kiril Simov, Nico Franz, Pavel Stoev, Terry Catapano, Donat Agosti, Guido Sautter, Robert A. Morris, and Lyubomir Penev. 2018. "OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system". *Journal of Biomedical Semantics* 9, no. 5 (). ISSN: 2041-1480. doi:10.1186/s13326-017-0174-5. https://jbiomedsem.biomedcentral.com/articles/

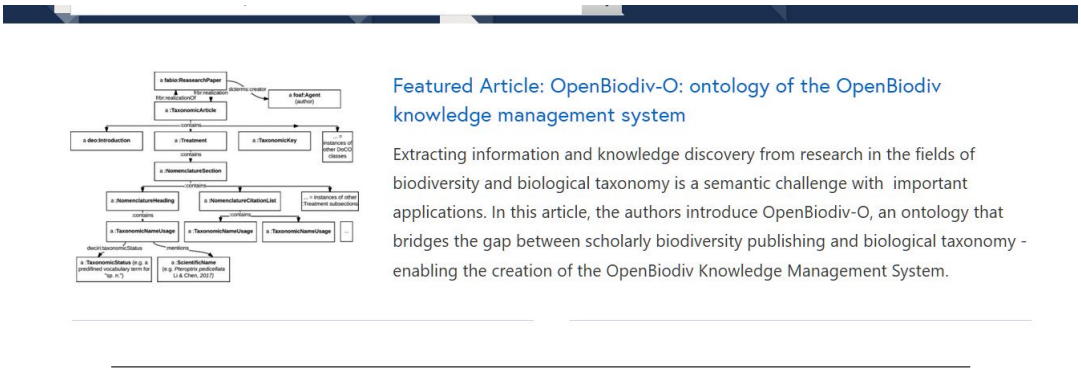


FIGURE 8.2: The OpenBiodiv-O article is featured on the main web-page of the Journal of Biomedical Semantics..

10.1186/s13326-017-0174-5 Indexed in WoS IF 1.6, Q3 SCOPUS, SJR 0.952. Unique citations by Michel et al., 2018, Page, 2018b, Page, 2018a.

A list of publications related to the dissertation follows. The listed articles have been published without exception in four international scientific journals: five articles in Research Ideas and Outcomes, one article in ZooKeys (WoS IF 1.079, Q3 SCOPUS, SJR 0.533), one article in Biodiversity Data Journal (WoS SCOPUS, SJR 0.465) and one article in Journal of Biomedical Semantics (WoS IF 1.6, Q3 SCOPUS, SJR 0.952). The total number of citations that have been accumulated for the candidate excluding self-citations (cross-citations) is at least 20. The citing articles are given in the list above. The total number of citations that have been accumulated including cross-citations and citations of work outside of the scope of the dissertation is at least 48 (Google Scholar).

[1] is an early version of the Introduction as well Chapter 1 and contains work towards Objective 2 (Architecture). The text of publications [2, 3, 5, 6, 7] are not a part of the text of the dissertation one-to-one but contain work towards Objective 5 (Workflows). The ideas presented in these publications have to large degree been incorporated in Chapter 5 whose backbone is formed by [4]; thus Objective 5 (Workflows) is achieved. [7] is published in the peer-reviewed journal ZooKeys with impact factor 1.031 (early 2018). [8] is the most important publication under this dissertation and was published in the high-impact Journal of Biomedical Semantics with impact factor 2.413 (early 2018). [8] makes up the content of Chapter 2 and is the main body of work fulfilling Objective 1 (Ontology). It was a featured article on the home-page of JBS (Fig. 8.2). Chapter 3 and Chapter 4 that form Objectives 3, 4, respectively are currently being prepared as manuscripts in international journals. Furthermore, the software library RDF4R described in Chapter 4 is being submitted to the open source repository rOpenSci[2].

Апробация на резултатите

Доклади пред научен семинар на ПНЗ

1. Доклад пред научен семинар на ИБЕИ на БАН на 26.10.2015 г. (“Публикуване, визуализация и разпространение на първични и геномни данни за биологичното разнообразие на основата на открита система за управление на информацията”).

[2]“We build software with a community of users and developers, and educate scientists about transparent research practices.” https://ropensci.org/

2. Доклад пред научен семинар в ИИКТ на БАН на 31.03.2016 г. (Open Biodiversity Knowledge Management System)

3. Доклад пред научен семинар на ИИКТ на БАН за 23.03.2018 г. (OpenBiodiv: a knowledge-based system of biodiversity information)

Доклади пред научно мероприятие в чужбина или пред международно научно мероприятие у нас

1. Доклад пред международния симпозиум EU BON в София на 23.03.2016 г. (The Data Publishing Toolkit at EU BON: Automated creation of data papers, data and text integrated publishing via the ARPHA Publishing Platform.)

2. Доклад по време на работната среща на BIG4 в Хавраники, Чехия на 03.06.2016 г. (Project Progress Report (OBKMS))

3. Доклад по време на работната среща на BIG4 в Хавраники, Чехия на 03.06.2016 г. (Modern Methods of Systematic Research and the BOLD Algorithm)

4. Уеб-базиран доклад (уебинар) пред международна аудитория в рамките на семинар на iDigBio на 16.07.2016 г. (Online direct import of specimen records from iDigBio infrastructure into taxonomic manuscripts)

5. Доклад по време на работната среща на BIG4 в Копенхаген на 14.10.2016 г. (Midterm Progress Report)

6. Доклад на международния симпозиум TDWG 2016 в Санта Клара де Сан Карлос от 5. до 9.12.2016 г. (Streamlining the Flow of Taxon Occurrence Data Between a Manuscript and Biological Databases)

7. Доклад на международния симпозиум TDWG 2016 в Санта Клара де Сан Карлос от 5. до 9.12.2016 г. (The Open Biodiversity Knowledge Management System: A Semantic Suite Running on top of the Biodiversity Knowledge Graph)

8. Доклад на международния симпозиум TDWG 2016 в Санта Клара де Сан Карлос от 5. до 9.12.2016 г. (Demonstrating the Prototype of the Open Biodiversity Knowledge Management System)

9. Доклад на международния симпозиум TDWG 2016 в Санта Клара де Сан Карлос от 5. до 9.12.2016 г. (Creation of Data Paper Manuscripts from Ecological Metadata Language (EML))

10. Уеб-базиран доклад пред междуродния семинар на работната група по семантични технология към Университета Вандербилт (Тенеси, САЩ) на 20.02.2017 г. (Open Biodiversity Knowledge Management System)

11. Доклад на европейската конференция на биосистематиците, BioSyst.eu 2016 на 15.08.2017 г. (The OpenBiodiv Knowledge System: The Future of Access to Biodiversity Knowledge)

12. Доклад на международния симпозиум TDWG 2017 в Отава, Канада от 1. до 6.10.2017 г. (OpenBiodiv Computer Demo: an Implementation of a Semantic System Running on top of the Biodiversity Knowledge Graph)

13. Доклад на международния симпозиум TDWG 2017 в Отава, Канада от 1. до 6.10.2017 г. (OpenBiodiv: an Implementation of a Semantic System Running on top of the Biodiversity Knowledge Graph)

14. Постер на международния симпозиум TDWG 2017 в Отава, Канада от 1. до 6.10.2017 г. (OpenBiodiv: an Implementation of a Semantic System Running on top of the Biodiversity Knowledge Graph)

15. Доклад по време на работната среща на BIG4 в Ла Палма, Испания от 30. окт. до 3 ноем. 2017 г. (Midterm Progress Report)

16. Доклад пред научен семинар на групата по биоинформатика (група Ронкуист) в Кралския природо-научен музей в Стокхолм на 29.11.2017 г.

Main scientific and applied contributions

In the course of the investigative effort, all six objectives have been achieved and the results have been published in international journals and have been presented at major conferences in Bulgaria and abroad. The most important contributions of the thesis are summarized as follows:

Декларация за оригиналност

Декларирам, че настоящата дисертация съдържа оригинални резултати, получени при проведени от мен научни изследвания, с подкрепата и съдействието на научния ми ръководител проф. Любомир Пенев и Издаделство Пенсофт, както и научния ми консултант доц. Кирил Симов и ИИКТ. Резултатите, които са получени, описани и/или публикувани от други учени, са надлежно и подробно цитирани в библиографията.

Настоящата дисертация не е прилагана за придобиване на научна степен в друго висше училище, университет или научен институт.

Виктор Сендеров

Acknowledgements

This research has been financed through the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 642241. My deep gratitude goes to the European Commission for enabling this wonderful opportunity!

I thank Prof. Lyubomir Penev and Prof. Kiril Simov for the valuable supervision. I also thank the staff and developers at Pensoft Publishers for the support in creating the platform and its popularization; in particular Prof. Pavel Stoev, Teodor Georgiev, Georgi Zhelezov, Iliyana Kuzmova, and Iva Kostadinova. Furthermore, I thank Pensoft's graphic designer, Slavena Peneva, for the help with creating the illustrations for this thesis and in presentations. Last but not least, Margarita Grudova and Elisaveta Taseva for providing valuable administrative support during the elaboration of the thesis.

I thank my colleagues from the Bulgarian Academy of Sciences (Institutes for Information and Communication Technologies and for Biodiviversity and Ecosystems Research) for their friendship and advice; in particular Prof. Galya Angelova, Prof. Boyko Georgiev, and Prof. Snejana Grozeva.

I thank my colleagues at the BIG4 training network for the feedback, friendship, and support. In particular Prof. Alexey Solovdnikov, but there are too many more names to mention.

I thank my international collaborators for their ideas, reviews, and collaboration on papers. In particular Prof. Nico Franz (Arizona State University), Dr. Daniel Mietchen (National Institutes of Health), Dr. Éamonn Ó Tuama (formerly at GBIF), and Prof. Bob Morris (emeritus UMASS).

I also thank everyone at Plazi for the co-ownership of the vision of the project; in particular, Dr. Donat Agosti, Terry Catapano, and Dr. Guido Sautter.

Last but not least, I would like to acknowledge Ontotext for building the GraphDB database and providing excellent support.

List of Abbreviations

AI	**A**rtificial **I**ntelligence
API	**A**pplication **P**rogramming **I**nterfaces
BDJ	**B**iodiveristy **D**ata **J**ournal
BIN	**B**arcode **I**dentification **N**umbers
BOLD	**B**arcode of **L**ife **D**ata **S**ystems
DBMS	**D**ata**D**ase **M**anagement **S**ystems
DEO	**D**iscourse **E**lement **O**ntology's
Darwin-SW	**Darwin S**emantic **W**eb
DwC	**D**ar**w**in **C**ore
DoCO	**Do**cument **C**omponent **O**ntology
DTD	**D**ocument **T**ype **D**efinition
FaBiO	**F**RBR-**a**ligned **Bi**bliographic **O**ntology
GBIF	**G**lobal **B**iodiversity **I**nformation **F**acility
GNA	**G**lobal **N**ames **A**rchitecture
ICZN	**I**nternational **C**ode of **Z**oological **N**omenclature
IICT	**I**nstitute of **I**nformation and **C**ommunication **T**echnology
KBMS	**K**nowledge **B**ase **M**anagement **S**ystem
LOD	**L**inked **O**pen **D**ata
OBKMS	**O**pen **B**iodiversity **K**nowledge **M**anagement **S**ystem
OECD	**O**rganization for **E**conomic **C**ooperation and **D**evelopment
OWL	**W**eb **O**ntology **L**anguage
RDF	**R**esource **D**escription **F**ormat
RDF4R	A an **R** package **for** working with **RDF**
RDFS	**RDF S**chema
RCC-5	**R**egion **C**onnection **C**alculus **5**
SKOS	**S**imple **K**nowledge **O**rganization **S**ystem
SPAR	**S**emantic **P**ublishing and **R**eferencing **O**ntologies
SPARQL	**SPARQL P**rotocol and **R**DF **Q**uery **L**anguage
TDWG	**A**Taxonomy **D**atabase **W**orking **G**roup
TNSS	**T**axonomic **N**omenclatural **S**tatus **T**erms
UI	**U**ser **I**nterface
XML	e**X**tensible **M**arkup **L**anguage

Bibliography

Agosti, D., C. Klingenberg, N. Johnson, C. Stephenson, and C. Catapano. 2007. *Why not let the computer save you time by reading the taxonomic papers for you?* Tech. rep. Zenodo. doi:10.5281/zenodo.15584.

Agosti, Donat. 2006. "Biodiversity data are out of local taxonomists' reach". *Nature* 439, no. 7075 (): 392–392. ISSN: 0028-0836, 1476-4687. doi:10.1038/439392a. http://www.nature.com/articles/439392a.

Allemang, Dean, and James A. Hendler. 2011. *Semantic Web for the working ontologist: effective modeling in RDFS and OWL*. 2nd ed. OCLC: ocn712780761. Waltham, MA: Morgan Kaufmann/Elsevier. ISBN: 978-0-12-385965-5.

Arriaga-Varela, Emmanuel, Matthias Seidel, Albert Deler-Hernández, Viktor Senderov, and Martin Fikácek. 2017. "A review of the Cercyon Leach (Coleoptera, Hydrophilidae, Sphaeridiinae) of the Greater Antilles". *ZooKeys* 681 (): 39–93. ISSN: 1313-2970, 1313-2989. doi:10.3897/zookeys.681.12522. https://zookeys.pensoft.net/articles.php?id=12522.

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. "DBpedia: A Nucleus for a Web of Open Data". In *The Semantic Web*, ed. by David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, 4825:722–735. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-540-76297-3 978-3-540-76298-0. doi:10.1007/978-3-540-76298-0_52. http://link.springer.com/10.1007/978-3-540-76298-0_52.

Bachman, Steven P, Eimear M Nic Lughadha, and Malin C Rivers. 2018. "Quantifying progress toward a conservation assessment for all plants". *Conservation Biology* 32 (3): 516–524.

Barrasa, Jesús. 2017. *RDF Triple Stores vs. Labeled Property Graphs: What's the Difference?* Published online. Visited on 01/11/2019. https://neo4j.com/blog/rdf-triple-store-vs-labeled-property-graph-difference/.

Baskauf, Steve, and Campbell O. Webb. 2016. "Darwin-SW: Darwin Core-based terms for expressing biodiversity data as RDF". *Semantic Web Journal* 7 (6): 629–643. doi:10.3233/SW-150203.

Beck, Kent, Mike Beedle, Arie van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, Ron Jeffries, Jon Kern, Brian Marick, Robert C. Martin, Steve Mellor, Ken Schwaber, Jeff Sutherland, and Dave Thomas. 2001. *Manifesto for Agile Software Development*. Published online. Visited on 01/11/2019. http://www.agilemanifesto.org/.

Beckett, David. 2014. *Redland RDF Libraries*. http://librdf.org/.

Berendsohn, Walter G. 1995. "The concept of" potential taxa" in databases". *Taxon*: 207–212. Visited on 07/22/2017. http://www.jstor.org/stable/1222443.

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The Semantic Web". *Scientific American* 284, no. 5 (): 34–43. ISSN: 0036-8733, visited on 04/13/2018. doi:10.1038/scientificamerican0501-34. http://www.nature.com/doifinder/10.1038/scientificamerican0501-34.

Bizer, Chris, and Richard Cyganiak. 2014. *RDF 1.1 TriG RDF Dataset Language W3C Recommendation 25 February 2014*. Visited on 07/17/2018. https://www.w3.org/TR/trig/.

Blomquist, HL. 1948. *The grasses of North Carolina*. Duke Universtiy Press.

Boettiger, Carl. 2018. *rdflib: A high level wrapper around the redland package for common rdf applications*. https://github.com/ropensci/rdflib.

Boettiger, Carl, Scott Chamberlain, Edmund Hart, and Karthik Ram. 2015. "Building software, building community: lessons from the rOpenSci project". *Journal of Open Research Software* 3 (1).

Bollacker, Kurt, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. "Freebase: a collaboratively created graph database for structuring human knowledge", 1247. ACM Press. ISBN: 978-1-60558-102-6, visited on 07/22/2018. doi:10.1145/1376616.1376746. http://portal.acm.org/citation.cfm?doid=1376616.1376746.

Brosens, Dimitri, Francois Vankerkhoven, David Ignace, Philippe Wegnez, Nico Noé, André Heughebaert, Jeannine Bortels, and Wouter Dekoninck. 2013. "FORMIDABEL: The Belgian Ants Database". *ZooKeys* 306 (): 59–70. ISSN: 1313-2970, 1313-2989, visited on 07/15/2018. doi:10.3897/zookeys.306.4898. http://zookeys.pensoft.net/articles.php?id=3172.

Burt, Trevor, and Ximo Mengual. 2017. "Origin and diversification of hoverflies: a revision of the genera Asarkina and Allobaccha–A BIG4 Consortium PhD project". *Research Ideas and Outcomes* 3:e19860.

Cardoso, Pedro, Pavel Stoev, Teodor Georgiev, Viktor Senderov, and Lyubomir Penev. 2016. "Species Conservation Profiles compliant with the IUCN Red List of Threatened Species". *Biodiversity Data Journal* 4 (): e10356. ISSN: 1314-2828, 1314-2836. doi:10.3897/BDJ.4.e10356. http://bdj.pensoft.net/articles.php?id=10356.

Catapano, Terence. 2010. "TaxPub: an extension of the NLM/NCBI journal publishing DTD for taxonomic descriptions". Published online. Visited on 07/11/2017. https://www.ncbi.nlm.nih.gov/books/NBK47081/.

Catapano, Terence, and Robert A. Morris. 2016. *Treatment Ontologies*. Visited on 08/09/2017. https://github.com/plazi/TreatmentOntologies/blob/master/treatment.owl.

Challenges in irreproducible research. 2010. Published online. Visited on 04/15/2018. https://www.nature.com/collections/prbfkwmwvz.

Chang, Winston. 2017. *R6: Classes with Reference Semantics*. Visited on 07/17/2018. https://cran.r-project.org/web/packages/R6/.

Chavan, Vishvas. 2013. "Cultural Change in Data Publishing Is Essential" [**inlang**en]. *BioScience* 63, no. 6 (): 419–420. ISSN: 00063568, 15253244, visited on 07/15/2018. doi:10.1525/bio.2013.63.6.3. https://academic.oup.com/bioscience/article-lookup/doi/10.1525/bio.2013.63.6.3.

Chavan, Vishwas, and Lyubomir Penev. 2011. "The data paper: a mechanism to incentivize data publishing in biodiversity science". *BMC Bioinformatics* 12 (Suppl 15): S2. ISSN: 1471-2105, visited on 02/13/2018. doi:10.1186/1471-2105-12-S15-S2. http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-S15-S2.

Chen, Mingmin, Shizhuo Yu, Nico Franz, Shawn Bowers, and others. 2014. "Euler/X: a toolkit for logic-based taxonomy integration". Published online, *arXiv preprint arXiv:1402.1992*. Visited on 08/11/2017. https://arxiv.org/abs/1402.1992.

Claerbout, Jon F., and Martin Karrenbach. 1992. "Electronic documents give reproducible research a new meaning", 601–604. Society of Exploration Geophysicists. Visited on 04/15/2018. doi:10.1190/1.1822162. http://library.seg.org/doi/abs/10.1190/1.1822162.

Constantin, Alexandru, Silvio Peroni, Steve Pettifer, David Shotton, and Fabio Vitali. 2016. "The Document Components Ontology (DoCO)". Ed. by Oscar Corcho. *Semantic Web* 7, no. 2 (): 167–181. ISSN: 22104968, 15700844, visited on 08/13/2017. doi:10.3233/SW-150177. http://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-150177.

Damova, Mariana, Atanas Kiryakov, Kiril Simov, and Svetoslav Petrov. 2010. "Mapping the central LOD ontologies to PROTON upper-level ontology". In *Proceedings of the 5th International Conference on Ontology Matching-Volume 689*, 61–72. CEUR-WS. org. Visited on 08/10/2017. http://dl.acm.org/citation.cfm?id=2878599.

DeVries, Peter. *Taxon Concept Ontology*. http://taxonconcept.org.

Deans, Andrew R., Matthew J. Yoder, and James P. Balhoff. 2012. "Time to change how we describe biodiversity". *Trends in Ecology & Evolution* 27, no. 2 (): 78–84. ISSN: 01695347, visited on 07/11/2017. doi:10.1016/j.tree.2011.11.007. http://linkinghub.elsevier.com/retrieve/pii/S0169534711003302.

Desmet, Peter, and Luc Brouilet. 2013. "Database of Vascular Plants of Canada (VASCAN): a community contributed taxonomic checklist of all vascular plants of Canada, Saint Pierre and Miquelon, and Greenland". *PhytoKeys* 25 (): 55–67. ISSN: 1314-2003, 1314-2011, visited on 07/15/2018. doi:10.3897/phytokeys.25.3100. http://www.pensoft.net/journals/phytokeys/article/3100/abstract/database-of-vascular-plants-of-canada-vascan-a-community-contributed-taxonomic-checklist-of-all-vascular-plants-of-canad.

Dmitriev, D.A., and M. Yoder. 2017. *NOMEN*. Published online. Visited on 07/22/2017. https://github.com/SpeciesFileGroup/nomen.

Egloff, Willi, David Patterson, Donat Agosti, and Gregor Hagedorn. 2014. "Open exchange of scientific knowledge and European copyright: The case of biodiversity information". *ZooKeys* 414 (): 109–135. ISSN: 1313-2970, 1313-2989, visited on 04/13/2018. doi:10.3897/zookeys.414.7717. http://zookeys.pensoft.net/articles.php?id=3830.

Fegraus, Eric H., Sandy Andelman, Matthew B. Jones, and Mark Schildhauer. 2005. "Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation" [**inlang**en]. *Bulletin of the Ecological Society of America* 86, no. 3 (): 158–168. ISSN: 0012-9623, visited on 07/15/2018. doi:10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2. http://doi.wiley.com/10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2.

Fielding, RT. 2000. *Architectural styles and the design of network-based software architectures. PhD Dissertation.* Irvine: Dept. of Information / Computer Science, University of California.

Filippova, Nina V., Ilya V. Filippov, Dmitry S. Schigel, Natalia V. Ivanova, and Maxim P. Shashkov. 2017. "Biodiversity informatics: global trends, national perspective and regional progress in Khanty-Mansi Autonomous Okrug". *Environmental Dynamics and Global Climate Change* 8, no. 2 (): 46–56. ISSN: 2541-9307, 2218-4422, visited on 03/15/2018. doi:10.17816/edgcc8246-56. http://journals.eco-vector.com/EDGCC/article/view/7080.

Franz, N.M., and R.K. Peet. 2009. "Perspectives: Towards a language for mapping relationships among taxonomic concepts". *Systematics and Biodiversity* 7, no. 1 (): 5–20. ISSN: 1477-2000, 1478-0933, visited on 07/22/2017. doi:10.1017/S147720000800282X. http://www.tandfonline.com/doi/abs/10.1017/S147720000800282X.

Franz, Nico M, and Beckett W Sterner. 2018. "To increase trust, change the social design behind aggregated biodiversity data". *Database* 2018.

Franz, Nico M., Mingmin Chen, Parisa Kianmajd, Shizhuo Yu, Shawn Bowers, Alan S. Weakley, and Bertram Ludäscher. 2016a. "Names are not good enough: Reasoning over taxonomic change in the Andropogon complex1". *Semantic Web* 7 (6): 645–667. Visited on 07/11/2017. http://content.iospress.com/articles/semantic-web/sw220.

Franz, Nico M., Naomi M. Pier, Deeann M. Reeder, Mingmin Chen, Shizhuo Yu, Parisa Kianmajd, Shawn Bowers, and Bertram Ludäscher. 2016b. "Two influential primate classifications logically aligned". *Systematic biology* 65 (4): 561–582. Visited on 07/11/2017. https://academic.oup.com/sysbio/article-abstract/65/4/561/1753624.

Franz, Nico, and Guanyang Zhang. 2017. "Three new species of entimine weevils in Early Miocene amber from the Dominican Republic (Coleoptera: Curculionidae)". *Biodiversity Data Journal* 5 (): e10469. ISSN: 1314-2828, 1314-2836, visited on 08/12/2017. doi:10.3897/BDJ.5.e10469. http://bdj.pensoft.net/articles.php?id=10469.

Franz, Nico, Chao Zhang, and Joohyung Lee. 2016c. "A logic approach to modeling nomenclatural change" (). Visited on 07/11/2017. doi:10.1101/058834. http://biorxiv.org/lookup/doi/10.1101/058834.

GBIF Secretariat. 2017. "GBIF Backbone Taxonomy. Checklist dataset". Published online. Visited on 06/30/2018. doi:10.15468/39omei. https://doi.org/10.15468/39omei.

Garnett, Stephen T., and Les Christidis. 2017. "Taxonomy anarachy hampers conservation". *Nature* 546 (). Visited on 08/12/2017. https://www.nature.com/polopoly_fs/1.22064!/menu/main/topColumns/topLeftColumn/pdf/546025a.pdf.

Godtsenhoven, Karen, van, Mikael Karstensen Elbaek, Barbara Sierman, Magchiel Bijsterbosch, Patrick Hochstenbach, Rosemary Russell, and Maurice Vanderfeesten. 2009. *Emerging Standards for Enhanced Publications and Repository Technology : Survey on Technology.* Amsterdam: Amsterdam University Press. ISBN: 978-90-8964-189-2, visited on 04/15/2018. doi:10.5117/9789089641892. http://dare.uva.nl/aup/nl/record/316870.

Grothendieck, G. 2018. *gsubfn: Utilities for Strings and Function Arguments*. Visited on 07/17/2018. https://cran.r-project.org/web/packages/gsubfn/index.html.

Gruber, Thomas R. 1993. "A translation approach to portable ontology specifications". *Knowledge Acquisition* 5, no. 2 (): 199–220. ISSN: 10428143, visited on 08/07/2017. doi:10.1006/knac.1993.1008. http://linkinghub.elsevier.com/retrieve/pii/S1042814383710083.

Guralnick, Robert, Tom Conlin, John Deck, Brian J. Stucky, and Nico Cellinese. 2014. "The Trouble with Triplets in Biodiversity Informatics: A Data-Driven Case against Current Identifier Practices" [**inlang**en], ed. by Damon P. Little. *PLoS ONE* 9, no. 12 (): e114069. ISSN: 1932-6203, visited on 04/15/2018. doi:10.1371/journal.pone.0114069. http://dx.plos.org/10.1371/journal.pone.0114069.

Gutt, Julian, David Barnes, Susanne J. Lockhart, and Anton van de Putte. 2013. "Antarctic macrobenthic communities: A compilation of circumpolar information". *Nature Conservation* 4 (): 1–13. ISSN: 1314-3301, 1314-6947, visited on 07/15/2018. doi:10.3897/natureconservation.4.4499. http://natureconservation.pensoft.net/articles.php?id=1342.

Harrington, Andrew. 2018. *Amortizing (not in Levitin)*. https://anh.cs.luc.edu/363/notes/06A_Amortizing.html.

Harris, Larry R., Jeffrey M. Hill, Dayton Marcott, and Timothy F. Rochford. 1993. Knowledge base management system. Pat. US5228116A, **patentfiled**July 1993. https://patents.google.com/patent/US5228116A/en.

Heath, Tom, and Christian Bizer. 2011. *Linked data: evolving the web into a global data space*. 1. ed. Synthesis lectures on the semantic web: theory and technology 1. OCLC: 732238828. San Rafael, Calif.: Morgan & Claypool. ISBN: 978-1-60845-430-3 978-1-60845-431-0.

Hong, Cui, Jocelyn Pender, and Brian Hedlund. 2018. *Explorer of Taxon Concepts*. https://docs.google.com/document/d/1F4vai5R7ygbUD3mopJxVh8ULQfa-x301l4tnTBRFVe4/edit?usp=sharing.

Huang, Fengqiong, James A Macklin, Hong Cui, Heather A Cole, and Lorena Endara. 2015. "OTO: Ontology Term Organizer" [**inlang**en]. *BMC Bioinformatics* 16, no. 1 (). ISSN: 1471-2105, visited on 08/11/2017. doi:10.1186/s12859-015-0488-1. http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0488-1.

International Commission on Zoological Nomenclature. 1999. *International Code of Zoological Nomenclature*. Fourth Edition. London, UK: The International Trust for Zoological Nomenclature. ISBN: 0 85301 006 4.

— . 2017. *The Official Registry of Zoological Nomenclature*. Publihsed online. Visited on 08/11/2017. http://zoobank.org/.

International code of nomenclature for algae, fungi and plants (Melbourne code): adopted by the Eighteenth International Botanical Congress Melbourne, Australia, July 2011. 2012. Regnum vegetabile v. 154. OCLC: ocn824722354. Königstein, Germany: Koeltz Scientific Books. ISBN: 978-3-87429-425-6.

Jansen, Michael Andrew, and Nico M. Franz. 2015. "Phylogenetic revision of Minyomerus Horn, 1876 sec. Jansen & Franz, 2015 (Coleoptera, Curculionidae) using taxonomic concept annotations and alignments". *ZooKeys* 528 (): 1–133. ISSN: 1313-2970, 1313-2989, visited on 08/12/2017. doi:10.3897/zookeys.528.6001. http://zookeys.pensoft.net/articles.php?id=6001.

Jarke, Matthias, Bernd Neumann, Yannis Vassiliou, and Wolfgang Wahlster. 1989. "KBMS Requirements of Knowledge-Based Systems". In *Foundations of Knowledge Base Management*, ed. by Michael L. Brodie, John Mylopoulos, Joachim W. Schmidt, Joachim W. Schmidt, and Constantino Thanos, 381–394. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-83399-1 978-3-642-83397-7, visited on 07/22/2018. doi:10.1007/978-3-642-83397-7_17. http://www.springerlink.com/index/10.1007/978-3-642-83397-7_17.

Jones, M, P Slaughter, J Ooms, C Boettiger, and S Chamberlain. 2016. "redland: RDF Library Bindings in R. R package version 1.0.17-9". doi:10.5063/F1VM496B. https://github.com/ropensci/redland-bindings/tree/master/R/redland.

Kennedy, Jessie B., Robert Kukla, and Trevor Paterson. 2005. "Scientific Names Are Ambiguous as Identifiers for Biological Taxa: Their Context and Definition Are Required for Accurate Data Integration". In *Data Integration in the Life Sciences: Second International Workshop, DILS 2005, San Diego, CA, USA, July 20-22, 2005. Proceedings*, ed. by Bertram Ludäscher and Louiqa Raschid, 80–95. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-540-31879-8. doi:10.1007/11530084_8.

Kissling, W Daniel, Ramona Walls, Anne Bowser, Matthew O Jones, Jens Kattge, Donat Agosti, Josep Amengual, Alberto Basset, Peter M Van Bodegom, Johannes HC Cornelissen, and others. 2018. "Towards global data products of Essential Biodiversity Variables on species traits". *Nature ecology & evolution*: 1.

Knuth, Donald Ervin. 1984. "Literate programming". *The Computer Journal* 27 (2): 97–111. Visited on 08/08/2017. https://academic.oup.com/comjnl/article-abstract/27/2/97/343244.

Koperski, M, M Sauer, W Braun, and S Gradstein. 2000. *Referenzliste der Moose Deutschlands*. Vol. 34. Schriftenreihe Vegetationsk.

Kraker, Peter, Derick Leony, Wolfgang Reinhardt, N.A. Gü, and nter Beham. 2011. "The case for an open science in technology enhanced learning". *International Journal of Technology Enhanced Learning* 3 (6): 643. ISSN: 1753-5255, 1753-5263, visited on 04/15/2018. doi:10.1504/IJTEL.2011.045454. http://www.inderscience.com/link.php?id=45454.

Kurtz, Jamie. 2013. "What is RESTful?" [**inlang**en]. In *ASP.NET MVC 4 and the Web API*, 9–21. Berkeley, CA: Apress. ISBN: 978-1-4302-4977-1 978-1-4302-4978-8, visited on 07/15/2018. doi:10.1007/978-1-4302-4978-8_2. http://link.springer.com/10.1007/978-1-4302-4978-8_2.

Kõljalg, Urmas, R. Henrik Nilsson, Kessy Abarenkov, Leho Tedersoo, Andy F. S. Taylor, Mohammad Bahram, Scott T. Bates, Thomas D. Bruns, Johan Bengtsson-Palme, Tony M. Callaghan, Brian Douglas, Tiia Drenkhan, Ursula Eberhardt, Margarita Dueñas, Tine Grebenc, Gareth W. Griffith, Martin Hartmann, Paul M. Kirk, Petr Kohout, Ellen Larsson, Björn D. Lindahl, Robert Lücking, María P. Martín, P. Brandon Matheny, Nhu H. Nguyen, Tuula Niskanen, Jane Oja, Kabir G. Peay, Ursula Peintner, Marko Peterson, Kadri Põldmaa, Lauri Saag, Irja Saar, Arthur Schüßler, James A. Scott, Carolina Senés, Matthew E. Smith, Ave Suija,

D. Lee Taylor, M. Teresa Telleria, Michael Weiss, and Karl-Henrik Larsson. 2013. "Towards a unified paradigm for sequence-based identification of fungi" [**inlang**en]. *Molecular Ecology* 22, no. 21 (): 5271–5277. ISSN: 09621083, visited on 08/12/2017. doi:10.1111/mec.12481. http://doi.wiley.com/10.1111/mec.12481.

Lepage, Denis, Gaurav Vaidya, and Robert Guralnick. 2014. "Avibase – a database system for managing and organizing taxonomic concepts". *ZooKeys* 420 (): 117–135. ISSN: 1313-2970, 1313-2989, visited on 08/13/2017. doi:10.3897/zookeys.420.7089. http://zookeys.pensoft.net/articles.php?id=3906.

Li, Diyan, Tiandong Che, Binlong Chen, Shilin Tian, Xuming Zhou, Guolong Zhang, Miao Li, Uma Gaur, Yan Li, Majing Luo, and others. 2017. "Genomic data for 78 chickens from 14 populations". *GigaScience* 6 (6): 1–5.

Lin, Qiang, Ying Qiu, Ruobo Gu, Meng Xu, Jia Li, Chao Bian, Huixian Zhang, Geng Qin, Yanhong Zhang, Wei Luo, Jieming Chen, Xinxin You, Mingjun Fan, Min Sun, Pao Xu, Byrappa Venkatesh, Junming Xu, Hongtuo Fu, and Qiong Shi. 2017. "Draft genome of the lined seahorse, Hippocampus erectus" [**inlang**en]. *GigaScience* 6, no. 6 (): 1–6. ISSN: 2047-217X, visited on 03/15/2018. doi:10.1093/gigascience/gix030. https://academic.oup.com/gigascience/article-lookup/doi/10.1093/gigascience/gix030.

Linnaeus, Carl von. 1758. *Systema naturae per regna tria naturae: secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. Volume 1.*

Mallet, James. 2001. "Species, concepts of". In *Encyclopedia of biodiversity*, 5:427–440. Published online. Visited on 07/11/2017. http://tarjomefa.com/wp-content/uploads/2016/02/4420-engilish.pdf.

Manktelow, Mariette. 2010. *History of taxonomy*. Visited on 07/22/2017. http://www.atbi.eu/summerschool/files/summerschool/Manktelow_Syllabus.pdf.

Mansinghka, Vikash, Richard Tibbetts, Jay Baxter, Pat Shafto, and Baxter Eaves. 2015. "BayesDB: A probabilistic programming system for querying the probable implications of data". ArXiv: 1512.05006, *arXiv:1512.05006 [cs]* (). Visited on 07/22/2018. http://arxiv.org/abs/1512.05006.

Marwick, Ben, and Suzanne Birch. 2017. *A Standard for the Scholarly Citation of Archaeological Data as an Incentive to Data Sharing*. Tech. rep. SocArXiv. Visited on 03/15/2018. doi:10.17605/OSF.IO/PY4HZ. https://osf.io/preprints/socarxiv/py4hz/.

Mathieu, Jérôme. 2018. "EGrowth: A global database on intraspecific body growth variability in earthworm". *Soil Biology and Biochemistry* 122:71–80.

Michel, Franck, Catherine Faron-Zucker, Sandrine Tercerie, and Gargominy Olivier. 2018. "Modelling Biodiversity Linked Data: Pragmatism May Narrow Future Opportunities". *Biodiversity Information Science and Standards* 2:e26235.

Michener, William K. 2006. "Meta-information concepts for ecological data management" [**inlang**en]. *Ecological Informatics* 1, no. 1 (): 3–7. ISSN: 15749541, visited on 07/15/2018. doi:10.1016/j.ecoinf.2005.08.004. http://linkinghub.elsevier.com/retrieve/pii/S157495410500004X.

Michener, William K., James W. Brunt, John J. Helly, Thomas B. Kirchner, and Susan G. Stafford. 1997. "NONGEOSPATIAL METADATA FOR THE ECOLOGICAL SCIENCES". *Ecological Applications* 7, no. 1 (): 330–342. ISSN: 1051-0761, visited on 02/14/2018. doi:10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2. http://doi.wiley.com/10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2.

Mietchen, Daniel. 2014. "The Transformative Nature of Transparency in Research Funding". *PLoS Biology* 12, no. 12 (): e1002027. ISSN: 1545-7885, visited on 04/15/2018. doi:10.1371/journal.pbio.1002027. http://dx.plos.org/10.1371/journal.pbio.1002027.

Milano, Filippo, Paolo Pantini, Stefano Mammola, and Marco Isaia. 2017. "LA CONSERVAZIONE DELL'ARANEOFAUNA IN ITALIA E IN EUROPA". *ATTI DELL'ACCADEMIA NAZIONALE ITALIANA DI ENTOMOLOGIA. RENDICONTI* 65:91–103.

Miles, Alistair, and Sean Bechofer. *SKOS Simple Knowledge Organization System RDF Schema*. Visited on 09/08/2017. https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html.

Miller, Jeremy, Torsten Dikow, Donat Agosti, Guido Sautter, Terry Catapano, Lyubomir Penev, Zhi-Qiang Zhang, Dean Pentcheff, Richard Pyle, Stan Blum, Cynthia Parr, Chris Freeland, Tom Garnett, Linda S Ford, Burgert Muller, Leo Smith, Ginger Strader, Teodor Georgiev, and Laurence Bénichou. 2012. "From taxonomic literature to cybertaxonomic content". *BMC Biology* 10 (1): 87. ISSN: 1741-7007, visited on 04/13/2018. doi:10.1186/1741-7007-10-87. http://bmcbiol.biomedcentral.com/articles/10.1186/1741-7007-10-87.

Miller, Jeremy, Donat Agosti, Lyubomir Penev, Guido Sautter, Teodor Georgiev, Terry Catapano, David Patterson, David King, Serrano Pereira, Rutger Vos, and Soraya Sierra. 2015. "Integrating and visualizing primary data from prospective and legacy taxonomic literature". *Biodiversity Data Journal* 3 (): e5063. ISSN: 1314-2828, 1314-2836, visited on 04/13/2018. doi:10.3897/BDJ.3.e5063. http://bdj.pensoft.net/articles.php?id=5063.

Mindell, David P., Brian L. Fisher, Peter Roopnarine, Jonathan Eisen, Georgina M. Mace, Roderic D. M. Page, and Richard L. Pyle. 2011. "Aggregating, Tagging and Integrating Biodiversity Research". Ed. by Sean A. Rands. *PLoS ONE* 6, no. 8 (): e19491. ISSN: 1932-6203, visited on 04/13/2018. doi:10.1371/journal.pone.0019491. http://dx.plos.org/10.1371/journal.pone.0019491.

Momtchev, Vassil, Deyan Peychev, Todor Primov, and Georgi Georgiev. 2009. "Expanding the pathway and interaction knowledge in linked life data". *Proc. of International Semantic Web Challenge*. Visited on 07/22/2017. http://challenge.semanticweb.org/documents/Linked%20Life%20Data-LLD%20semantic%20web%20challenge%202009.pdf.

Morris, Paul J., Robert A. Morris, and Zhimin Wang. *Taxonomic Nomenclatural Status Terms*. Visited on 12/01/2018. https://github.com/pensoft/OpenBiodiv/blob/master/ontology/contrib/taxonomic_nomenclatural_status_terms.owl.

Mulsant, E. 1866. "Monographie des Coccinellides". *Mém. L'Acad. Imp. Lyon, Cl. Sci., LMém. L'Acad. Imp. Lyon.* 15:1–112.

Neo4J Developers. 2012. *Neo4J Graph NoSQL Database*. Published online. http://neo4j.com.

Nguyen, Nhung T. H., Axel J. Soto, Georgios Kontonatsios, Riza Batista-Navarro, and Sophia Ananiadou. 2017. "Constructing a biodiversity terminological inventory" [**inlang**en], ed. by Robert Guralnick. *PLOS ONE* 12, no. 4 (): e0175277. ISSN: 1932-6203, visited on 08/12/2017. doi:10.1371/journal.pone.0175277. http://dx.plos.org/10.1371/journal.pone.0175277.

Obitko, Marek. 2007. "Translations between ontologies in multi-agent systems". Ph. D. Dissertation, , Czech Technical University, Faculty of Electrical Engineering.

Ontotext. 2018. *GraphDB 8.6*. Published online. Visited on 07/17/2018. http://graphdb.ontotext.com/.

Open Biodiversity Knowledge Management System (OBKMS). 2014. Visited on 04/13/2018. http://adm.pro-ibiosphere.eu/getatt.php?filename=oo_4749.pdf.

Ordynets, Alexander, Anton Savchenko, Alexander Akulov, Eugene Yurchenko, Vera Malysheva, Urmas Kõljalg, Josef Vlasák, Karl-Henrik Larsson, and Ewald Langer. 2017. "Aphyllophoroid fungi in insular woodlands of eastern Ukraine". *Biodiversity Data Journal* 5 (): e22426. ISSN: 1314-2828, 1314-2836, visited on 03/15/2018. doi:10.3897/BDJ.5.e22426. https://bdj.pensoft.net/articles.php?id=22426.

Page, R. D. M. 2008. "Biodiversity informatics: the challenge of linking data and the role of shared identifiers". *Briefings in Bioinformatics* 9, no. 5 (): 345–354. ISSN: 1467-5463, 1477-4054, visited on 04/13/2018. doi:10.1093/bib/bbn022. https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbn022.

— . 2015. *Putting some bite into the Bouchout Declaration*. Published online. http://iphylo.blogspot.bg/2015/05/putting-some-bite-into-bouchout.html.

— . 2014. *The vision thing - it's all about the links*. Published online. http://iphylo.blogspot.bg/2014/06/the-vision-thing-it-all-about-links.html.

Page, Roderic D. M. 2016a. "DNA barcoding and taxonomy: dark taxa and dark texts" [**inlang**en]. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371, no. 1702 (): 20150334. ISSN: 0962-8436, 1471-2970, visited on 08/07/2017. doi:10.1098/rstb.2015.0334. http://rstb.royalsocietypublishing.org/lookup/doi/10.1098/rstb.2015.0334.

— . 2006. "Taxonomic names, metadata, and the Semantic Web". *Biodiversity Informatics* 3, no. 0 (). ISSN: 15469735, visited on 04/13/2018. doi:10.17161/bi.v3i0.25. https://journals.ku.edu/index.php/jbi/article/view/25.

— . 2012. *The GBIF classification is broken — how do we fix it?* Visited on 08/12/2017. http://iphylo.blogspot.bg/2012/05/gbif-classification-is-broken-how-do-we.html.

Page, Roderic DM. 2018a. "Liberating links between datasets using lightweight data publishing: an example using plant names and the taxonomic literature". Published online, *bioRxiv*: 343996.

— . 2018b. "Ozymandias: A biodiversity knowledge graph". Published online, *bioRxiv*: 485854.

Page, Roderic. 2016b. "Towards a biodiversity knowledge graph". *Research Ideas and Outcomes* 2 (): e8767. ISSN: 2367-7163, visited on 07/23/2017. doi:10.3897/rio.2.e8767. http://rio.pensoft.net/articles.php?id=8767.

Parr, Cynthia S., Robert Guralnick, Nico Cellinese, and Roderic D.M. Page. 2012. "Evolutionary informatics: unifying knowledge about the diversity of life" [**inlang**en]. *Trends in Ecology & Evolution* 27, no. 2 (): 94–103. ISSN: 01695347, visited on 04/13/2018. doi:10.1016/j.tree.2011.11.001. http://linkinghub.elsevier.com/retrieve/pii/S0169534711003247.

Patterson, D.J., J. Cooper, P.M. Kirk, R.L. Pyle, and D.P. Remsen. 2010. "Names are key to the big new biology". *Trends in Ecology & Evolution* 25, no. 12 (): 686–691. ISSN: 01695347, visited on 07/11/2017. doi:10.1016/j.tree.2010.09.004. http://linkinghub.elsevier.com/retrieve/pii/S0169534710002181.

Patterson, David J., David Remsen, William A. Marino, and Cathy Norton. 2006. "Taxonomic Indexing—Extending the Role of Taxonomy". Ed. by Rod Page. *Systematic Biology* 55, no. 3 (): 367–373. ISSN: 1076-836X, 1063-5157, visited on 11/16/2018. doi:10.1080/10635150500541680. https://academic.oup.com/sysbio/article/55/3/367/1667279.

Pellissier Tanon, Thomas, Denny Vrandečić, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. 2016. "From Freebase to Wikidata: The Great Migration", 1419–1428. ACM Press. ISBN: 978-1-4503-4143-1, visited on 07/22/2018. doi:10.1145/2872427.2874809. http://dl.acm.org/citation.cfm?doid=2872427.2874809.

Penev, Lyubomir, Teodor Georgiev, Peter Geshev, Seyhan Demirov, Viktor Senderov, Iliyana Kuzmova, Iva Kostadinova, Slavena Peneva, and Pavel Stoev. 2017a. "ARPHA-BioDiv: A toolbox for scholarly publication and dissemination of biodiversity data based on the ARPHA Publishing Platform". *Research Ideas and Outcomes* 3, no. e13088 (). ISSN: 2367-7163. doi:10.3897/rio.3.e13088. http://riojournal.com/articles.php?id=13088.

Penev, Lyubomir, W. John Kress, Sandra Knapp, De-Zhu Li, and Susanne Renner. 2010a. "Fast, linked, and open – the future of taxonomic publishing for plants: launching the journal PhytoKeys". *PhytoKeys* 1, no. 0 (). ISSN: 1314-2003, 1314-2011, visited on 07/22/2017. doi:10.3897/phytokeys.1.642. http://www.pensoft.net/journal_home_page.php?journal_id=3&page=article&type=show&article_id=642&abstract=1.

Penev, Lyubomir, Terence Catapano, Donat Agosti, Teodor Georgiev, Guido Sautter, and Pavel Stoev. 2012. "Implementation of TaxPub, an NLM DTD extension for domain-specific markup in taxonomy, from the experience of a biodiversity publisher". Published online. Visited on 07/11/2017. https://www.ncbi.nlm.nih.gov/books/NBK100351/.

Penev, Lyubomir, Daniel Mietchen, Vishwas Chavan, Gregor Hagedorn, David Remsen, Vincent Smith, and David Shotton. 2016. "Pensoft Data Publishing Policies And Guidelines For Biodiversity Data" (). Visited on 07/15/2018. doi:10.5281/zenodo.56660. https://zenodo.org/record/56660.

Penev, Lyubomir, Donat Agosti, Teodor Georgiev, Terry Catapano, Jeremy Miller, Vladimir Blagoderov, David Roberts, Vincent Smith, Irina Brake, Simon Ryrcroft, Ben Scott, Norman Johnson, Robert Morris, Guido Sautter, Vishwas Chavan, Tim Robertson, David Remsen, Pavel Stoev, Cynthia Parr, Sandra Knapp, W. John Kress, Frederic Thompson, and Terry Erwin. 2010b. "Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples". *ZooKeys* 50 (): 1–16. ISSN: 1313-2970, 1313-2989, visited on 04/15/2018. doi:10.3897/zookeys.50.538. http://zookeys.pensoft.net/articles.php?id=2215.

Penev, Lyubomir, Daniel Mietchen, Vishwas Chavan, Gregor Hagedorn, Vincent Smith, David Shotton, Éamonn Ó Tuama, Viktor Senderov, Teodor Georgiev, Pavel Stoev, Quentin Groom, David Remsen, and Scott Edmunds. 2017b. "Strategies and guidelines for scholarly publishing of biodiversity data". *Research Ideas and Outcomes* 3, no. e12431 (). ISSN: 2367-7163. doi:10.3897/rio.3.e12431. http://riojournal.com/articles.php?id=12431.

Penev, Lyubomir, Christopher Lyal, Anna Weitzman, David Morse, David King, Guido Sautter, Teodor Georgiev, Robert Morris, Terry Catapano, and Donat Agosti. 2011. "XML schemas and mark-up practices of taxonomic literature". *ZooKeys* 150 (): 89–116. ISSN: 1313-2970, 1313-2989, visited on 07/04/2018. doi:10.3897/zookeys.150.2213. http://zookeys.pensoft.net/articles.php?id=3038.

Peroni, Silvio. 2015. *Example of use of DoCO #2*. Visited on 08/12/2017. doi:10.6084/m9.figshare.1513733. http://figshare.com/articles/Example_of_use_of_DoCO_2/1513733.

— . 2014. "The semantic publishing and referencing ontologies". In *Semantic Web Technologies and Legal Scholarly Publishing*, 1st ed., 15:121–193. Springer. Visited on 07/22/2017.

Peroni, Silvio, and David Shotton. 2012. "FaBiO and CiTO: Ontologies for describing bibliographic resources and citations". *Web Semantics: Science, Services and Agents on the World Wide Web* 17 (): 33–43. ISSN: 15708268, visited on 08/13/2017. doi:10.1016/j.websem.2012.08.001. http://linkinghub.elsevier.com/retrieve/pii/S1570826812000790.

Pierrat, Benjamin, Thomas Saucède, Alain Festeau, and Bruno David. 2012. "Antarctic, Sub-Antarctic and cold temperate echinoid database". *ZooKeys* 204 (): 47–52. ISSN: 1313-2970, 1313-2989, visited on 07/15/2018. doi:10.3897/zookeys.204.3134. http://zookeys.pensoft.net/articles.php?id=2881.

Platnick, Norman I. 2001. "From cladograms to classifications: The road to DePhylocode". *The Systematics Association*. Visited on 08/07/2017. https://www.researchgate.net/profile/Norman_Platnick/publication/254335935_From_Cladograms_to_Classifications_The_Road_to_DePhylocode/links/0c96053b54b985b178000000.pdf.

Poorani, J., and Roger Booth. 2016. "Harmonia manillana (Mulsant), a new addition to Indian Coccinellidae, with changes in synonymy". *Biodiversity Data Journal* 4 (): e8030. ISSN: 1314-2828, 1314-2836, visited on 08/13/2017. doi:10.3897/BDJ.4.e8030. http://bdj.pensoft.net/articles.php?id=8030.

Pyle, Richard. 2016a. *Taxonomic name usage files*. Visited on 08/13/2017. http://lists.tdwg.org/pipermail/tdwg-content/2016-April/003582.html.

— . 2016b. "Towards a Global Names Architecture: The future of indexing scientific names". *ZooKeys* 550 (): 261–281. ISSN: 1313-2970, 1313-2989, visited on 08/12/2017. doi:10.3897/zookeys.550.10009. http://zookeys.pensoft.net/articles.php?id=6241.

Pyron, Robert Alexander. 2018. "A 21st Century Vision for Neotropical Snake Systematics". *Revista Latinoamericana de Herpetología* 1 (1): 58–62.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Published online, Vienna, Austria. https://www.R-project.org/.

RDF Working Group. 2014. *Resource Description Framework (RDF)*. Visited on 07/17/2018.

RDF4J development team. 2017. *The RDF4J Server REST API*. Visited on 07/17/2018.

Ratnasingham, Sujeevan, and Paul D. N. Hebert. 2013. "A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System" [**inlang**en], ed. by Diego Fontaneto. *PLoS ONE* 8, no. 7 (): e66213. ISSN: 1932-6203, visited on 07/22/2017. doi:10.1371/journal.pone.0066213. http://dx.plos.org/10.1371/journal.pone.0066213.

Rebholz-Schuhmann, Dietrich, Harald Kirsch, and Francisco Couto. 2005. "Facts from text—is text mining ready to deliver?" *PLoS biology* 3 (2): e65. Visited on 07/22/2017. http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0030065.

Remsen, David. 2016. "The use and limits of scientific names in biological informatics". *ZooKeys* 550 (): 207–223. ISSN: 1313-2970, 1313-2989, visited on 07/11/2017. doi:10.3897/zookeys.550.9546. http://zookeys.pensoft.net/articles.php?id=6234.

Sarah Faulwetter, Evangelos Pafilis, Lucia Fanini, Nicolas Bailly, Donat Agosti, Christos Arvanitidis, Laura Boicenco, Terry Catapano, Simon Claus, Stefanie Dekeyzer, Teodor Georgiev, Aglaia Legaki, Dimitra Mavraki, Anastasis Oulas, Gabriella Papastefanou, Lyubomir Penev, Guido Sautter, Dmitry Schigel, Viktor Senderov, Adrian Teaca, and Marilena Tsompanou. 2016. "EMODnet Workshop on mechanisms and guidelines to mobilise historical data into biogeographic databases". *Research Ideas and Outcomes* 2, no. e10445 (). ISSN: 2367-7163. doi:10.3897/rio.2.e10445. http://rio.pensoft.net/articles.php?id=10445.

Senderov, Viktor, and Lyubomir Penev. 2016. "The Open Biodiversity Knowledge Management System in Scholarly Publishing". *Research Ideas and Outcomes* 2, no. e7757 (). ISSN: 2367-7163. doi:10.3897/rio.2.e7757. http://rio.pensoft.net/articles.php?id=7757.

Senderov, Viktor, Teodor Georgiev, and Lyubomir Penev. 2016. "Online direct import of specimen records into manuscripts and automatic creation of data papers from biological databases". *Research Ideas and Outcomes* 2 (): e10617. ISSN: 2367-7163. doi:10.3897/rio.2.e10617. http://rio.pensoft.net/articles.php?id=10617.

Senderov, Viktor, Kiril Simov, Nico Franz, Pavel Stoev, Terry Catapano, Donat Agosti, Guido Sautter, Robert A. Morris, and Lyubomir Penev. 2018. "OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system". *Journal of Biomedical Semantics* 9, no. 5 (). ISSN: 2041-1480. doi:10.1186/s13326-017-0174-5. https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-017-0174-5.

Senderov, Viktor, Nico M. Franz, and Kiril Simov. 2017. *OpenBiodiv Ontology and Guide*. Published online. Visited on 08/09/2017. http://openbiodiv.net/ontology.

Shao, Kwang-tsao, Jack Lin, Chung-Han Wu, Hsin-Ming Yeh, and Tun-Yuan Cheng. 2012. "A dataset from bottom trawl survey around Taiwan". *ZooKeys* 198 (): 103–109. ISSN: 1313-2970, 1313-2989, visited on 07/15/2018. doi:10.3897/zookeys.198.3032. http://zookeys.pensoft.net/articles.php?id=2824.

Shotton, David. 2009. "Semantic publishing: the coming revolution in scientific journal publishing". *Learned Publishing* 22, no. 2 (): 85–94. ISSN: 09531513, visited on 04/15/2018. doi:10.1087/2009202. http://doi.wiley.com/10.1087/2009202.

Singhal, Amit. 2012. *Introducing the knowledge graph: things, not strings*. Published online. Visited on 01/11/2019. https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html.

Smith, Vincent, Teodor Georgiev, Pavel Stoev, Jordan Biserkov, Jeremy Miller, Laurence Livermore, Edward Baker, Daniel Mietchen, Thomas Couvreur, Gregory Mueller, Torsten Dikow, Kristofer M. Helgen, Jiři Frank, Donat Agosti, David Roberts, and Lyubomir Penev. 2013. "Beyond dead trees: integrating the scientific process in the Biodiversity Data Journal". *Biodiversity Data Journal* 1 (): e995. ISSN: 1314-2828, 1314-2836, visited on 08/10/2017. doi:10.3897/BDJ.1.e995. http://bdj.pensoft.net/articles.php?id=995.

Sokal, Robert R. 1963. "The Principles and Practice of Numerical Taxonomy". *Taxon* 12, no. 5 (): 190. ISSN: 00400262, visited on 08/12/2017. doi:10.2307/1217562. http://www.jstor.org/stable/1217562?origin=crossref.

Staab, Steffen, and Rudi Studer, eds. 2009. *Handbook on Ontologies*. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-540-70999-2 978-3-540-92673-3, visited on 08/07/2017. doi:10.1007/978-3-540-92673-3. http://link.springer.com/10.1007/978-3-540-92673-3.

Sterner, Beckett, and Nico M. Franz. 2017. "Taxonomy for Humans or Computers? Cognitive Pragmatics for Big Data". *Biological Theory* 12, no. 2 (): 99–111. ISSN: 1555-5542, 1555-5550, visited on 07/11/2017. doi:10.1007/s13752-017-0259-5. http://link.springer.com/10.1007/s13752-017-0259-5.

Taxonomic Names and Concepts Interest Group. 2006. *Taxonomic Concept Transfer Schema (TCS), version 1.01*. Published online. Visited on 01/11/2019. http://www.tdwg.org/standards/117.

Taylor, Gary, D. Austin Andy, Jennings John, Purcell Matthew, and Wheeler Greg. 2010. "Casuarinicola, a new genus of jumping plant lice (Hemiptera: Triozidae) from Casuarina (Casuarinaceae)". *Zootaxa*, no. 2601: 1–27.

Tennant, Jonathan P, Jonathan M Dugan, Daniel Graziotin, Damien C Jacques, François Waldner, Daniel Mietchen, Yehia Elkhatib, Lauren B Collister, Christina K Pikas, Tom Crick, and others. 2017. "A multi-disciplinary perspective on emergent and future innovations in peer review". *F1000Research* 6.

The Apache Software Foundation. 2013. *Apache Lucene - Query Parser Syntax*. https://lucene.apache.org/core/2_9_4/queryparsersyntax.html.

The Bouchout Declaration for Open Biodiversity Knowledge Management. 2014. Published online. Visited on 01/11/2019. http://www.bouchoutdeclaration.org/fileadmin/Dateien/PDF/Bouchout_Declaration_EN.pdf.

The W3C SPARQL Working Group, ed. 2013. *SPARQL 1.1 Overview W3C Recommendation 21 March 2013*. Visited on 07/17/2018. https://www.w3.org/TR/2013/REC-sparql11-overview-20130321/.

Thorpe, Stephen. 2013. "Casuarinicola australis Taylor, 2010 (Hemiptera: Triozidae), newly recorded from New Zealand". *Biodiversity Data Journal* 1 (): e953. ISSN: 1314-2828, 1314-2836, visited on 08/11/2017. doi:10.3897/BDJ.1.e953. http://bdj.pensoft.net/articles.php?id=953.

Tillett, Barbara. 2003. *A conceptual model for the Bibliographic Universe*. Vol. 25. 5. Technicalities. Visited on 08/08/2017. http://www.loc.gov/cds/downloads/FRBR.PDF.

Tng, David, Deborah Apgaua, Mason Campbell, Casey Cox, Darren Crayn, Françoise Ishida, Michael Liddell, Michael Seager, and Susan Laurance. 2016. "Vegetation and floristics of a lowland tropical rainforest in northeast Australia". *Biodiversity Data Journal* 4 (): e7599. ISSN: 1314-2828, 1314-2836, visited on 07/15/2018. doi:10.3897/BDJ.4.e7599. http://bdj.pensoft.net/articles.php?id=7599.

Trontelj, P, and C Fiser. 2009. "Cryptic species should not be trivialized". *Systematics and Biodiversity*, no. 7: 1–23.

Tzitzikas, Yannis, Carlo Allocca, Chryssoula Bekiari, Yannis Marketakis, Pavlos Fafalios, Martin Doerr, Nikos Minadakis, Theodore Patkos, and Leonardo Candela. 2013. "Integrating heterogeneous and distributed information about marine species through a top level ontology". In *Research Conference on Metadata and Semantic Research*, 289–301. Springer. Visited on 07/22/2017. http://link.springer.com/chapter/10.1007/978\-3\-319\-03437\-9_29.

University of Copenhagen, University of Turku, Institute for Systematic Zoology and Evolutionary Biology, Zoologisches Forschungsmuseum Alexander Koenig, Naturhistorisches Museum Wien, Pensoft Publishers company, ERA7 Bioinformatics, Swedish Museum of Natural History, and Decuria IT company. 2014. *BIG4 - Biosystematics, Informatics and Genetics of the big 4 insect groups: training tomorrow's researchers and entrepreneurs*. Published online. Visited on 01/11/2019. http://big4-project.eu.

Vrandečić, Denny, and Markus Krötzsch. 2014. "Wikidata: a free collaborative knowledgebase". *Communications of the ACM* 57, no. 10 (): 78–85. ISSN: 00010782, visited on 07/22/2018. doi:10.1145/2629489. http://dl.acm.org/citation.cfm?doid=2661061.2629489.

Walls, Ramona L., John Deck, Robert Guralnick, Steve Baskauf, Reed Beaman, Stanley Blum, Shawn Bowers, Pier Luigi Buttigieg, Neil Davies, Dag Endresen, Maria Alejandra Gandolfo, Robert Hanner, Alyssa Janning, Leonard Krishtalka, Andréa Matsunaga, Peter Midford, Norman Morrison, Eamonn O. Tuama, Mark Schildhauer, Barry Smith, Brian J. Stucky, Andrea Thomer, John Wieczorek, Jamie Whitacre, and John Wooley. 2014. "Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies" [**inlang**en], ed. by Vladimir B. Bajic. *PLoS ONE* 9, no. 3 (): e89606. ISSN: 1932-6203, visited on 10/30/2017. doi:10.1371/journal.pone.0089606. http://dx.plos.org/10.1371/journal.pone.0089606.

Was ist Open Science? Published online. Visited on 01/11/2019. http://opensciencеasap.org/open-science/.

What is GBIF? Published online. Visited on 08/12/2017. http://www.gbif.org/what-is-gbif.

Wickham, Hadley. 2015. *Advanced R*. The R series. OCLC: ocn881664644. Boca Raton, FL: CRC Press. ISBN: 978-1-4665-8696-3.

— . 2017. *httr: Tools for Working with URLs and HTTP*. Published online. Visited on 07/17/2018. https://cran.r-project.org/web/packages/httr/.

Wickham, Hadley, Jim Hester, and Winston Chang. 2018a. *devtools: Tools to Make Developing R Packages Easier*. Visited on 07/17/2018. https://cran.r-project.org/web/packages/devtools/index.html.

Wickham, Hadley, James Hester, and Jeroen Ooms. 2018b. *xml2: Parse XML*. Published online. Visited on 07/17/2018. https://cran.r-project.org/web/packages/xml2/index.html.

Wieczorek, John, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Renato Giovanni, Tim Robertson, and David Vieglais. 2012. "Darwin Core: An Evolving Community-Developed Biodiversity Data Standard". Ed. by Indra Neil Sarkar. *PLoS ONE* 7, no. 1 (): e29715. ISSN: 1932-6203, visited on 07/22/2017. doi:10.1371/journal.pone.0029715. http://dx.plos.org/10.1371/journal.pone.0029715.

Williams, Antony J., Lee Harland, Paul Groth, Stephen Pettifer, Christine Chichester, Egon L. Willighagen, Chris T. Evelo, Niklas Blomberg, Gerhard Ecker, Carole Goble, and Barend Mons. 2012. "Open PHACTS: semantic interoperability for drug discovery". *Drug Discovery Today* 17, **numbers** 21-22 (): 1188–1198. ISSN: 13596446, visited on 07/22/2017. doi:10.1016/j.drudis.2012.05.016. http://linkinghub.elsevier.com/retrieve/pii/S1359644612001936.

Witteveen, Joeri. 2015. "Naming and contingency: the type method of biological taxonomy". *Biology & Philosophy* 30, no. 4 (): 569–586. ISSN: 0169-3867, 1572-8404, visited on 08/13/2017. doi:10.1007/s10539-014-9459-6. http://link.springer.com/10.1007/s10539-014-9459-6.

Wolfram|Alpha, Making the wolrd's knowledge computable. Published online. Wolfram Alpha LLC. Visited on 06/10/2018. https://www.wolframalpha.com/.

pro-iBiosphere project final report. 2014. Published online. Visited on 04/13/2018. http://adm.pro-ibiosphere.eu/getatt.php?filename=oo_4751.pdf.

pro-iBiosphere. 2013. *Competency Questions for RDF Treatments*. Visited on 08/12/2017. http://wiki.pro-ibiosphere.eu/wiki/Competency_Questions_for_RDF_Treatments.

pro-iBiosphere. Visited on 08/12/2017. http://wiki.pro-ibiosphere.eu/.

Филиппова, НВ, ИВ Филиппов, ДС Щигель, НВ Иванова, and МП Шашков. 2017. "Информатика биоразнообразия: мировые тенденции, состояние дел в России и развитие направления в Ханты-Мансийском Автономном Округе". *Динамика окружающей среды и глобальные изменения климата* 8 (2): 46–56.

Шашков, МП, ИФ Чадин, and НВ Иванова. 2017. "Методические рекомендации по стандартизации данных для публикации через глобальный портал GBIF. ORG и подготовке статьи о данных". *Труды Кольского научного центра РАН*, no. 6-5 (8).

Шашков, МП, and НВ Иванова. 2018. "Стандарты и веб-инструменты для публикации данных через глобальные порталы по биоразнообразию". *Доклады Международной конференции "Математическая биология и биоинформатика"* 7:e98. doi:10.17537/icmbb18.55.