publications

MDPI

Article

OpenBiodiv: A Knowledge Graph for Literature-Extracted Linked Open Data in Biodiversity Science

Lyubomir Penev [1,2,*], Mariya Dimitrova [1,3], Viktor Senderov [4], Georgi Zhelezov [1], Teodor Georgiev [1], Pavel Stoev [1,5] and Kiril Simov [3]

1 Pensoft Publishers, Prof. Georgi Zlatarski Street 12, 1700 Sofia, Bulgaria; m.dimitrova@pensoft.net (M.D.); gzhelezov@pensoft.net (G.Z.); preprint@pensoft.net (T.G.); projects@pensoft.net (P.S.)
2 Institute of Biodiversity and Ecosystem Research, Bulgarian Academy of Sciences, 2 Gagarin Street, 1113 Sofia, Bulgaria
3 Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Acad. G. Bonchev St., Block 25A, 1113 Sofia, Bulgaria; kivs@bultreebank.org
4 Swedish Museum of Natural History, Frescativägen 40, 114 18 Stockholm, Sweden; vsenderov@gmail.com
5 National Museum of Natural History, 1 Tsar Osvoboditel Blvd, 1000 Sofia, Bulgaria
* Correspondence: penev@pensoft.net

Received: 23 April 2019; Accepted: 24 May 2019; Published: 29 May 2019

Abstract: Hundreds of years of biodiversity research have resulted in the accumulation of a substantial pool of communal knowledge; however, most of it is stored in silos isolated from each other, such as published articles or monographs. The need for a system to store and manage collective biodiversity knowledge in a community-agreed and interoperable open format has evolved into the concept of the Open Biodiversity Knowledge Management System (OBKMS). This paper presents OpenBiodiv: An OBKMS that utilizes semantic publishing workflows, text and data mining, common standards, ontology modelling and graph database technologies to establish a robust infrastructure for managing biodiversity knowledge. It is presented as a Linked Open Dataset generated from scientific literature. OpenBiodiv encompasses data extracted from more than 5000 scholarly articles published by Pensoft and many more taxonomic treatments extracted by Plazi from journals of other publishers. The data from both sources are converted to Resource Description Framework (RDF) and integrated in a graph database using the OpenBiodiv-O ontology and an RDF version of the Global Biodiversity Information Facility (GBIF) taxonomic backbone. Through the application of semantic technologies, the project showcases the value of open publishing of Findable, Accessible, Interoperable, Reusable (FAIR) data towards the establishment of open science practices in the biodiversity domain.

Keywords: open science; biodiversity; biodiversity informatics; knowledge management system; semantic publishing; Linked Open Data; Semantic Web; ontology

1. Introduction

1.1. Background

Biodiversity science studies and describes the diversity of living organisms on Earth. It is an interdisciplinary field that encompasses knowledge from multiple domains: Taxonomy, genomics, biogeography, ecology, phylogenetics and others. Developing mechanisms for storage and management of such diverse and rich information is of particular importance for biology and several other areas of research and practical activities [1–3].

Attempts to create a system and standards for integrating biodiversity knowledge have existed since the establishment of the Biodiversity Informatics Standards organization in 1985, then called Taxonomy Database Working Group (TDWG) [4]. Other major initiatives, like the Global Biodiversity Information Facility (GBIF), were started with the aim towards "an international mechanism... to make biodiversity data and information accessible worldwide" [5]. Inspired by the concept of Open Science, the idea for an Open Biodiversity Knowledge Management System (OBKMS) was developed during the EU-funded pro-iBiosphere project, which aimed to put forward an information system for biodiversity science [6,7]. The first steps in this direction were taken with the sanctioning of the Bouchout Declaration, which outlined several fundamental principles of open data in the biodiversity domain [8]. Alongside policy recommendations for removing the legal barriers that hinder the open sharing of biodiversity knowledge, the Bouchout Declaration, pro-iBiosphere and several other initiatives recommended linking biodiversity knowledge via unique, stable and resolvable identifiers and relevant infrastructure for storing and managing data [1,6,9–11]. Now signed by over 95 organizations and more than 200 researchers from around the world, the Bouchout Declaration is a communal pledge to the promotion of open science and the application of the Findable, Accessible, Interoperable, Reusable (FAIR) principles for biodiversity data [8,12].

The integration of published research narrative and data into a single information space is crucial for the consolidation of biodiversity knowledge but is still limited by insufficient adoption of shared standards for data, as well as of routine workflows for data publishing and/or semantic markup of the narrative [13]. The development of an OBKMS is essential for taxonomy and all sciences that use Latin scientific names for organisms (Linnaean names). Information related to Latin names, as a rule, is locked in discrete articles and data repositories isolated from one another [14,15]. Besides the need to identify organisms, Linnaean names also have to reflect taxonomic viewpoints about classification of organisms within the taxonomic hierarchy [15]. Integration of taxonomic data, however, is complicated by the ambiguous nature of Linnaean names due to cases of synonymy, homonymy, the principles of priority (in case of several names for a given taxon, the oldest name published in accordance with the rules of the respective biological code of nomenclature is considered valid), revisions of taxonomic classifications and polysemy [14,16,17]. As a result, Linnaean names are subject to change. Tracking these changes through time and having the ability to store and query taxonomic information, often assembled within taxon-specific descriptions called taxonomic treatments, is vital for biodiversity science and the related fields of research and practice [2,13,15]. In contrast to Linnaean names, globally unique identifiers (GUIDs) can be used to unambiguously identify entities by linking objects and Web resources that describe them [12]. The use of globally unique and stable identifiers goes far beyond the basic need to reconcile Linnaean names. The complexity of biodiversity data requires many data elements that represent concepts and entities, such as collections, specimens, tissues, samples, publications, taxon name usages (TNUs) within publications, literature references and individual images, amongst many others, to be unambiguously identified and linked in a single data space [17]. This highlights the need for a system that can ensure interoperability between various data types scattered through different sources within a single environment, based on community accepted standards like stable unique identifiers, Resource Description Framework (RDF) and Linked Open Data. The application of these practices would facilitate a "greater alignment between data and expertise" [15] connecting the various users, stakeholders, scientific domains and areas of practical implementation, such as nature conservation and climate change mitigation [3].

1.2. Working Examples of Biodiversity Data Platforms and Knowledge Management Systems

The principles of linked open data (LOD) [18] present a way to organize knowledge management systems for the Web. According to them, each managed knowledge resource receives a stable, unique and resolvable identifier on the Web. Provision of open access to linked datasets on the Web increases data discoverability and its impact on different communities. As early as 1993 [19], knowledge management systems (KMS) were defined as consisting of a knowledge store and a logic layer of rules,

which enables the inference of new facts. The use of common standards like the RDF [18] within KMS allows the linking of datasets from different domains and also facilitates the integration of an ontological layer, which grants the KMS a capability for logical reasoning. Examples of KMS integrating knowledge from different domains and conforming to the LOD principles are WikiData [19] and DBPedia [20]. Despite the focus on LOD to implement KMS solutions, there are many highly successful examples of large biodiversity data aggregation platforms using more conventional technologies, such as relational or wide-column databases (reviewed by Bingham et al. [21]). Most data platforms typically focus on a limited range of data types (e.g., occurrence or genomic data) or only treat important taxa or model organisms of high interest. Notable examples include the Global Biodiversity Information Facility (GBIF) for species occurrence data [4]; Catalogue of Life (CoL) for taxonomic classification [21]; the International Nucleotide Sequence Database Collaboration (INSDC), comprising GenBank, DDBJ and ENA databases, for gene sequence data [22]; the Barcode of Life Data System (BOLD) for sequences of certain barcode genes plus data about voucher specimens [23] and, on a smaller scale, databases based on particular taxa: Avibase [24], the world bird database, or Diptera [25], a data hub for the order Diptera (flies). Column-based solutions in each of these platforms allow the implementation of data constraints and normalization, which effectively contribute to curation of high-quality data. Thus, relational databases and wide-column stores are key to the first stages of data management and use.

The presence of multiple sources of occurrence, collection, morphology, taxonomy and genomic data, however, significantly depreciates the potential of the relational database model for storing and using numerous data types within a single data platform [3]. Thus, column-based database models do not provide the means for a centralized system for storing, indexing and accessing all available biodiversity knowledge. The next step in the data management process is providing comprehensive knowledge access, as outlined in the "Global Biodiversity Informatics Outlook", published by GBIF [26]. The document highlights indexing and structured data formats as some of the key elements towards this goal, as well as the challenges before "delivering access to all published biodiversity knowledge" [26]. The above-mentioned platforms for biodiversity data each have their own framework for storing and managing data, and some even use different taxonomic backbones. LOD-based solutions for biodiversity data attempt to integrate different aspects of biodiversity by taking advantage of common standards and data formats. Mapping entities from different sources via universal identifiers allows LOD frameworks to generate a network of data, thus complementing column-based solutions. The shared vision of an "inter-connected digital knowledge base" [26] suggests KMS as the obvious choice for improving global access to biodiversity data.

As a synthetic dataset, the OpenBiodiv LOD relies on the coherence of data compiled into the dataset. In many cases, these data have been stored and managed using a column-based solution, which reiterates that OpenBiodiv is complementary to column-based databases and not an alternative to them. For instance, the RDF triples encoding GBIF's taxonomic backbone are a direct transformation of the data stored within GBIF's wide-column database containing the backbone.

There are two pioneering attempts for creating knowledge management systems in the biodiversity domain, often named "biodiversity graphs"—OpenBiodiv [6] and Ozymandias [27]. They serve as discovery tools for biodiversity data available on the Web.

The knowledge graph Ozymandias was recently launched by Roderic Page for the Australian fauna [27]. By using shared identifiers, Ozymandias links taxonomic classifications, publications, article metadata and images by integrating data from the Atlas of Living Australia (ALA), the Australian Faunal Directory (AFD), CrossRef, the Biodiversity Literature Repository (BLR), Biostor, ORCID and Wikispecies. Hence, it constitutes an important proof of concept for employing semantic-based knowledge graphs to discover connected entities from the biodiversity domain.

The present paper describes the rationale, concept, infrastructure and underlying data of OpenBiodiv—the first LOD-based OBKMS, which integrates knowledge extracted from biodiversity publications and a taxonomic backbone tree used by GBIF. The development of OpenBiodiv began

in 2015 [7], a proof of concept being launched and presented during the annual conference of the Biodiversity Information Standards (TDWG) in 2016 [28].

2. Materials and Methods

Conceptual modelling of the biodiversity knowledge extracted from publications is provided in the OpenBiodiv-O ontology, described in detail in [29]. It introduced several new ontological classes and relationships to integrate already existing biodiversity-specific ontologies (e.g., DarwinCore (DwC) [30]) with scholarly publishing ontologies (e.g., Semantic Publishing and Referencing Ontologies (SPAR) [31]).

RDF data in OpenBiodiv was extracted from more than 5000 articles published in several open access journals by Pensoft Publishers, namely *ZooKeys*, *PhytoKeys*, *MycoKeys*, *Journal of Orthoptera Research* and the *Biodiversity Data Journal*, as well as from Plazi's Treatment Bank, which contains digitized legacy taxonomic treatments from more than 20,000 articles scattered through approximately 80 academic journals [1]. Scientific names used within this large group of texts were mapped to GBIF's taxonomic backbone, which has been converted into RDF and integrated in OpenBiodiv.

Pensoft's journal articles are published in eXtensible Markup Language (XML) according to the TaxPub XML schema [32] and are semantically tagged and enhanced with publishing and biodiversity metadata [33–35]. Likewise, Plazi's taxonomic treatments are extracted from the published literature and converted to XML using the TaxonX schema [1,34,36].

Information extraction and the subsequent conversion of XML into RDF triples is performed with the help of the open source R packages RDF4R and ROpenBio, created as part of the project and openly available at [37,38], respectively. GBIF's taxonomic backbone [39] was downloaded and transformed to RDF using a custom PHP script. Generation of RDF triples is controlled via scripts from the OpenBiodiv base package, available at [40]. The RDF statements were uploaded to a graph database repository (Ontotext's GraphDB) [41], which is accessible at [42].

Retrieving knowledge from the OpenBiodiv database is possible via either the SPARQL endpoint [42] or a free-text search functionality at the website [43]. The website, aimed at users with little or no prior knowledge of SPARQL, offers full text search via the indexing library Apache Lucene®, which enables fuzzy matching of searched entities.

3. Results

3.1. System Architecture and Data Model

The realization of OpenBiodiv as an Open Biodiversity Knowledge Management System was done through the creation of a semantic database, which contains a Linked Open Dataset based on the ontology OpenBiodiv-O [28], a codebase for automatic transformation of literature into RDF statements, a website [43] providing a frontend to the database and a SPARQL endpoint [42] (Figure 1).

3.2. The OpenBiodiv-O Ontology

OpenBiodiv-O serves as a semantic framework for the database. It can also be used on its own to understand the intersection between the domains of biodiversity and scholarly publishing [29]. OpenBiodiv-O helps to represent the structure of taxonomic articles via the introduction of new ontological classes and relationships (Table 1). Some of them include: Treatment, Nomenclature Section, Taxonomic Name Usage and Taxonomic Concept [29]. These resource types are unique to taxonomic literature and, to our best knowledge, have not previously been modelled in an ontology before.

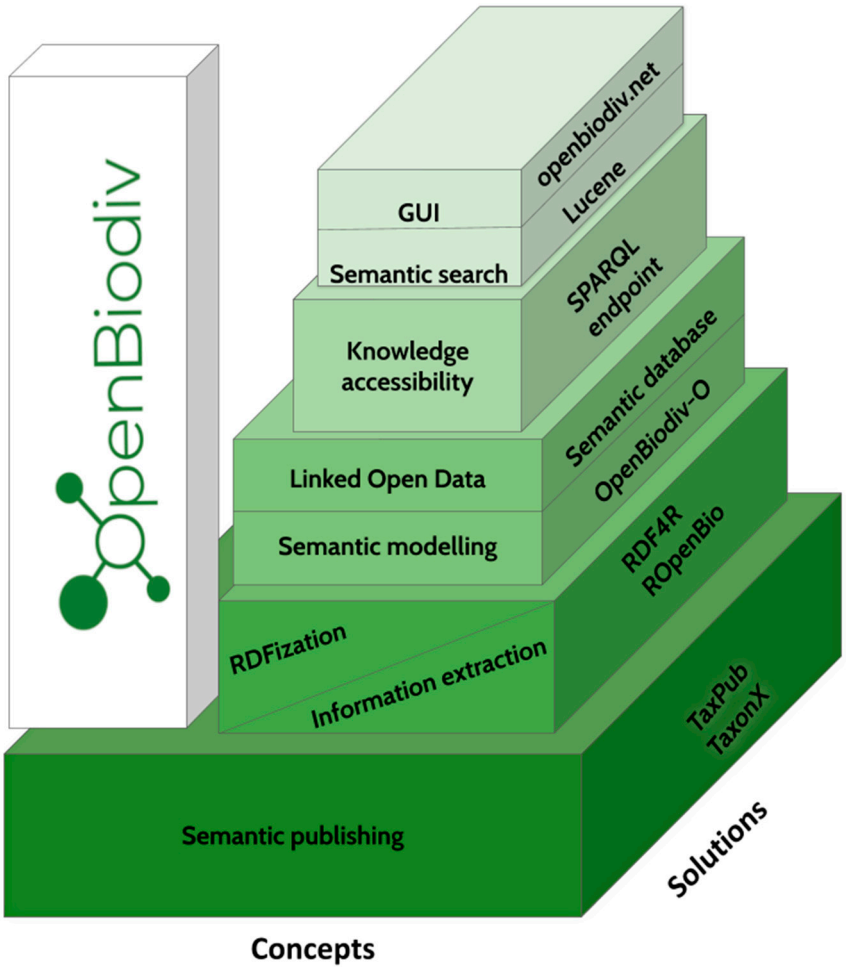

Figure 1. Architectural model of OpenBiodiv.

Table 1. Entity types extracted from scholarly literature and represented within OpenBiodiv via the OpenBiodiv-O ontology. The sources for resource types are noted via the ontology URLs. Adapted from Senderov et al. [29].

Entity Type	Comment	Ontology Source
Journal Article	A scientific article	http://purl.org/spar/fabio
Article Title	The title of an article	http://purl.org/dc/elements/1.1/
Digital Object Identifier (DOI)	The DOI of an article	http://prismstandard.org/namespaces/basic/2.0/
Introduction	The Introduction section of an article	http://www.sparontologies.net/ontologies/deo
Author Name	The name of an article author	http://xmlns.com/foaf/0.1/
Treatment	Section of a taxonomic article	http://openbiodiv.net/ontology
Nomenclature Section	Subsection of Treatment	http://openbiodiv.net/ontology
Nomenclature Citation List	List of citations of related concepts	http://openbiodiv.net/ontology
Materials Examined	List of examined specimens	http://openbiodiv.net/ontology
Biology Section	Subsection of Treatment	http://openbiodiv.net/ontology
Description Section	Subsection of Treatment	http://openbiodiv.net/ontology
Taxonomic Key	Section with an identification key	http://openbiodiv.net/ontology
Taxonomic Checklist	Section with a list of taxa for a region	http://openbiodiv.net/ontology
Taxonomic Name Usage	Mention of a taxonomic name	http://openbiodiv.net/ontology
Taxonomic Concept	Contextualized use of a taxonomic name, including a literature source	http://openbiodiv.net/ontology

In addition, OpenBiodiv-O is used to infer new knowledge based on known relationships between different entity types. The ontology was imported into the GraphDB repository instance holding the OpenBiodiv dataset and, as a result, a total of nearly 452 million new semantic triples were generated, thus representing a 2.63 inference ratio.

3.3. The OpenBiodiv Knowledge Base

The OpenBiodiv database was designed to reflect key biodiversity data and publishing metadata extracted from literature. Some of the extracted entity types, including types unique to the OpenBiodiv ontology, are listed in Table 1.

Pensoft's semantic publishing workflow [33,34] and Plazi's text and data extraction and publishing workflow [1] (Figure 2), previously reviewed by Penev et al. [34], enabled the effective transformation of XML-marked up text into semantic statements. Standardized tag sets from TaxPub [32] and TaxonX [44] provided mechanisms to map extracted atomized data from the XML documents to objects from the ontology. The open accessibility of the packages RDF4R [37] and ROpenBio [38] ensures the reproducibility of all processes and enables the modification of code to model other XML schemas or add new types of extracted data to the knowledge base in the future.

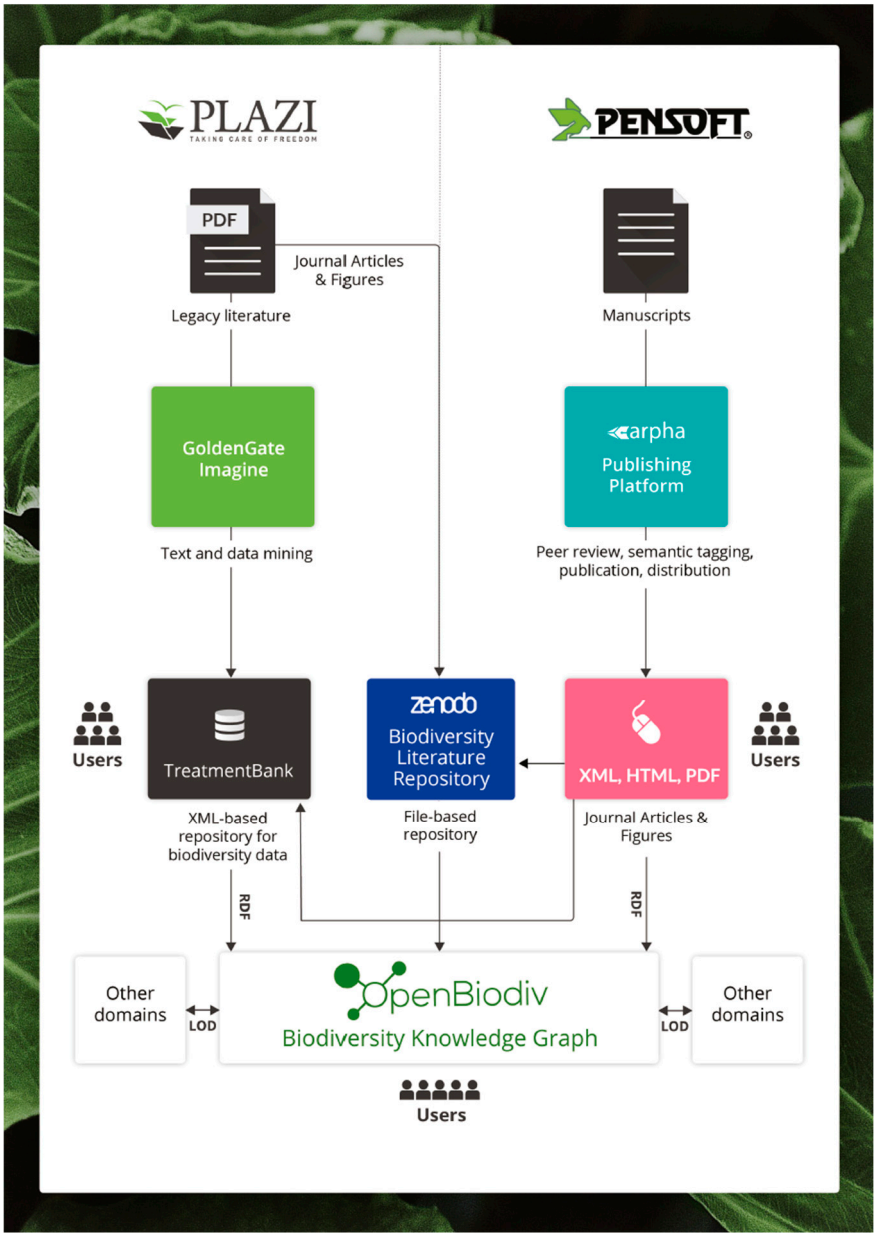

Figure 2. Data extraction and Resource Description Framework (RDF) conversion workflows from prospectively published literature (Pensoft) and legacy publications (Plazi) that feed into triples uploaded to OpenBiodiv [45].

The transformation of structured text from scholarly literature into RDF triples contributed to the creation of the LOD dataset, at present containing 729,263,110 statements (see Table 2 for some other statistics, such as number of articles, authors and scientific names, compared to those in Ozymandias). By transforming scientific name usages from research articles into distinct Web resources, OpenBiodiv establishes a name checklist of its own, which helps to elucidate the overlap between published scientific names and synthetic checklists, like GBIF's taxonomic backbone.

Table 2. Basic features and statistic figures of OpenBiodiv and Ozymandias.

	OpenBiodiv	**Ozymandias**
Year of launch of a prototype	2016 (TDWG 2016) [27]	2018 (Ebbe Nielsen Challenge 2018) [26]
Ontology used	OpenBiodiv-O	schema.org, TAXREF, TDWG LSID
Main resource types	Journal, Article, Article metadata, Article sections, Figure legend, Taxonomic treatment, Taxonomic treatment subsections, Taxonomic concept, Taxonomic name usage	Journal, Article, Article metadata, Figure, Taxonomic name, Taxonomic concept
Taxonomic classification	GBIF	Atlas of Living Australia (ALA)
Reconciliation of authors and publications	In progress	CrossRef, ORCID, Wikispecies, Biostor
Additional identifier cross-linking	-	Wikidata, GBIF
Figures	Figure legends from taxonomic publications and treatments	Biodiversity Literature Repository
	Number of entity instances	
Journals	35	6210
Journal articles	24,212	68,217
Authors	38,800	32,548
Scientific names	6,704,000 (extracted from literature) 5,798,686 (imported from GBIF)	444,222

3.4. Semantic Search

The development of diverse use cases and possible user scenarios for OpenBiodiv was the keystone for the development of the knowledge base from a theoretical concept into a practically orientated Open Biodiversity Knowledge Management System. A wide range of questions, concerning both publications metadata and content, can be answered by querying the database. Table 3 lists some example questions, along with the relevant user groups that could be interested in them and their potential usefulness. Apart from scientists, who can use the system to find organism descriptions and related organisms, various institutional organizations, such as natural history museums or others concerned with conservation, research funding or education, can also benefit from OpenBiodiv.

While the SPARQL endpoint to the database [42] allows users to write their own queries, certain questions can be answered by typing a simple search term in the website [43]. The latter does not require any prior knowledge of SPARQL or an understanding of the hierarchy of ontological classes and relationships within the graph. The first question, "Are there any articles mentioning the scientific name X and how many are there?" (Table 3), can be answered through the frontend Web portal by entering a specific scientific name into the search box. Upon recognition of the search entity type, a relevant SPARQL query is executed to retrieve RDF statements from the knowledge graph. The results are then formatted and displayed in a separate results page.

Table 3. Examples of competency questions and user stories that can be addressed by OpenBiodiv.

Question	Target User Groups	Value
Are there any articles mentioning the scientific name X and how many are there?	Taxonomists, ecologists and practitioners	Evaluation of the current state of research of taxon name X
Which specimens from a certain collection have been used/cited in publications and which are these publications?	Natural history collection managers and administrators; taxonomists	Tracking usage of collection material of particular value (holotypes, type series, extinct taxa, other material)
Which taxon treatments (or other general article sections) mention both scientific name X and Y?	Taxonomists, ecologists	Identification of taxa that are potentially related
How many articles about taxon X has a given researcher written in the past 10 years?	Research institutions, funding bodies, biodiversity researchers	Evaluation of a scientist's research impact and expertise (e.g., during the grant proposal writing process)
How many articles about a taxon X are published over a certain period of time?	Research institutions, funding bodies, ecological organizations, biodiversity researchers	Identification of poorly known species to evaluate the need for funding and conducting research; facilitation of literature discovery and research

4. Discussion

4.1. OpenBiodiv as a Major Step towards FAIR Data and Open Science

Providing open access to the scientific knowledge about Earth's biodiversity is the overarching goal for the OBKMS [6,8]. The establishment of an OBKMS during the development of OpenBiodiv was achieved through the use of stable identifiers for resources, semantic modelling of the biodiversity publishing domain, text and data mining and integration of multiple isolated content silos into a single Linked Open Dataset. Through the transformation of disparate statements within research articles and their subsections into connected semantic statements, the project unlocks knowledge hidden within these sources, including such behind a paywall or other access barriers [1], and opens it to the world of Linked Open Data.

Generating the synthetic Linked Open Dataset was significantly facilitated by the use of machine-readable formats in Pensoft's and Plazi's [1,34] publication workflows based on XML. XML is a markup format that helps to confer additional meaning and structure to areas within the text of a scientific article. This enrichment allows identifying scientific names, authors, institutions and taxonomic sections within research articles. Pensoft's and Plazi's adoption of the XML schemas TaxPub [32] and TaxonX [44] contribute to a standardized mechanism for these entities in scholarly literature and constitutes an accomplishment of one of the fundamental principles of OBKMS: The use of agreed vocabularies and standards [6].

In the process of development of OpenBiodiv-O [29], several other established ontologies and vocabularies were reused in accordance with the principles advocated by the proponents of the Semantic Web [6,46]. Reusing vocabularies facilitates linking of entities within the OpenBiodiv LOD to entities from external data stores that use these vocabularies. Both primary sources, Pensoft's articles and Plazi's extracted taxonomic treatments, together with the OpenBiodiv dataset itself, are openly available on the Web. The dataset is available as RDF triples, which is a machine-readable, non-proprietary format, as well as an established standard of the World Wide Web Consortium for universal resource identification [18]. Finally, the incorporation of multiple ontologies within OpenBiodiv-O, as well as the mapping of scientific names to terms from GBIF's taxonomic backbone, makes the data linked to other datasets on the Web.

From the start of its conceptual design to its implementation, OpenBiodiv constitutes an open science project for the biodiversity domain. The system utilizes data embedded in the published scientific literature through data liberation, extraction and transformation into a standardized

machine-readable format, which further enhances the data by making it FAIR [12]. The semantic workflow of OpenBiodiv is openly accessible through the project documentation and program code and can be applied to any domain. Thus, OpenBiodiv represents an example for conducting open and reproducible research and showcases an important use case for a valuable utilization via a knowledge graph of both prospectively published content and that extracted from the legacy literature.

OpenBiodiv is a direct outcome of the process started within the pro-iBiosphere project [6] with elaboration of the concept of Open Biodiversity Knowledge Management (OBKM) and ending with the publication of the Bouchout Declaration [8]. The declaration appealed to institutions, individuals, non-governmental organizations and the general public to join efforts and forces towards "free and open access to data and information about biodiversity by people and computers" [8]. By establishing an infrastructure and workflow to liberate information from literature, linking it "using agreed vocabularies" and making it accessible [8], OpenBiodiv is one of the very first practical implementations of the OBKM concept. Utilizing Linked Open Data for storing and linking biodiversity knowledge could stimulate a synergy of multiple similar projects for biodiversity.

4.2. Comparison with Other Biodiversity Knowledge Graphs

OpenBiodiv is a specialized knowledge management system, focused solely on scientific biodiversity information. In contrast, other knowledge management systems, like DBPedia and WikiData aim to establish cross-domain knowledge graphs [19,20]. The knowledge statements in OpenBiodiv are automatically generated from the narrative of research texts, together with the appropriate provenance information.

On the other hand, knowledge statements in WikiData are crowd-sourced, and provenance is entered by the Wikipedia volunteers. DBPedia is based on an automatic extraction of Wikipedia knowledge, which in itself is created by volunteers. Thus, unlike errors in WikiData and DBPedia, the potential errors in OpenBiodiv are machine errors.

Another important aspect of knowledge graphs is federation—the ability to link knowledge from several graphs via the matching of globally unique identifiers for the same resources stored in different graphs. For example, digital object identifiers (DOIs) uniquely and globally identify articles and can be used to cross-link article information across different knowledge graphs. To achieve decentralized building of his Ozymandias biodiversity knowledge graph (see Table 2 for a comparison between OpenBiodiv and Ozymandias), Page has used global identifiers, such as DOIs and ORCIDs, to map entities from different data sources into the knowledge base [27]. For instance, DOIs are used for publication and author reconciliation between the Australian Faunal Directory, CrossRef, ORCID and Wikispecies. Hence, Ozymandias constitutes a mashup of different data sources to harvest additional knowledge.

Similarly, OpenBiodiv links entities originating from Pensoft, Plazi and GBIF. However, its main focus is discovering entity relationships within primary sources (see Table 3 for comparison between the two knowledge graphs). By modelling individual article sections using multiple publishing ontology terms, OpenBiodiv allows the tracking of the contextual usage of taxonomic names within these sections. This is particularly important for the nomenclature section of taxonomic treatments, which provide information about the relationship between names. In addition, OpenBiodiv maps local identifiers for taxonomic names used in an article to the identifiers from GBIF's backbone taxonomy.

In contrast to this model, Ozymandias uses vocabulary terms from schema.org, TAXREF and TDWG LSID, which, Page argues, is preferable to the use of multiple domain-specific ontologies because it offers a simpler ontology model, is widely adopted by the search engines of Google, Microsoft and Yahoo and allows embedding of data from the knowledge graph into websites [27]. While convenient for these purposes, vocabularies used in Ozymandias do not offer the granularity of domain-specific ontologies for conceptually modelling something as specific as taxonomic articles (see Figure 3 for a comparison of the two conceptual models). In the case of Ozymandias, journal articles are represented as a "single, monolithic entity" [27] and individual sections are not distinguished (Figure 3B).

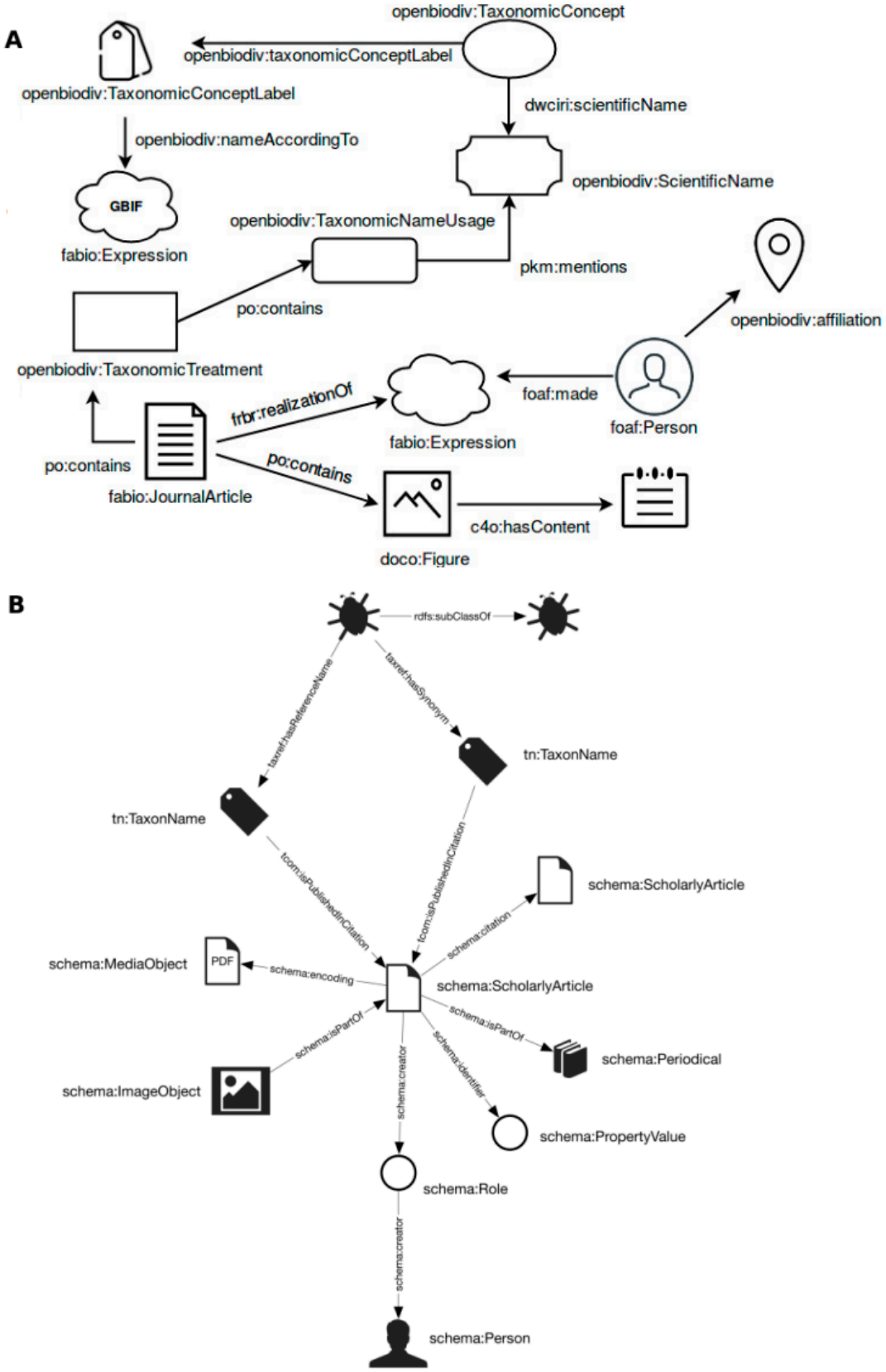

Figure 3. Conceptual modelling of a taxonomic article in OpenBiodiv (**A**) and Ozymandias (**B**) (re-used from [27]).

In Ozymandias the most granular units of information are figures and articles. Thus, the usage of names within distinct article sections cannot be tracked [27]. Article figures in Ozymandias are harvested from BLR using the DOI of the article, and a regular expression is used for the retrieval of taxon-specific images, a process which can lead to the incorrect exclusion of some figures. In-depth modelling of taxonomic articles and their subsections, including linking figure legends to taxonomic names in treatments, allows the accurate extraction of specific image data from the XML of the article itself. This type of linking is only possible through text mining of the taxonomic treatment section. A more complex but also more fine-grained ontology, like OpenBiodiv-O, provides the semantic architecture to represent these relationships between entities in taxonomic literature, thus enhancing the potential of the biodiversity graph for conducting taxonomic meta-analyses. The recent introduction of new formatting guidelines for the "Material Citations" section by the *European Journal of Taxonomy* and Pensoft's journals [47,48] would enable an even more detailed markup of taxonomic articles in the future, which, in combination with changes to the ontology, would enrich the knowledge stored in OpenBiodiv.

4.3. Limitations and Future Directions

OpenBiodiv successfully demonstrated that creating a knowledge graph from information extracted and formalized from scholarly biodiversity literature is a realistic and promising perspective. However, being in the beta stage of its development, the project still has some limitations. They can be divided into four groups related to: (1) Reusing and sharing identifiers with other sources of biodiversity knowledge, (2) user interaction with the knowledge graph, (3) type of mapped entities from scholarly literature, and (4) disambiguation. The following subsections explore the current limitations and provide future directions for their resolution.

4.3.1. Reusing and Sharing Identifiers with Other Sources of Biodiversity Knowledge

Reusing of identifiers is one of the key principles of LOD and knowledge graphs. During the process of triple generation in OpenBiodiv, identifiers are either extracted from XML, retrieved from the knowledge base if they already exist there or created. Processing of an XML file involves extraction of all relevant nodes according to the used schema and establishing identifiers for them. Hence, each article, subsection of an article and taxonomic name usage receives a separate identifier. In the case of treatments from Plazi, the treatment identifier given by Plazi is the only extracted and reused identifier. Similarly, figures with their own DOIs from Pensoft's journal articles are given new identifiers instead of reusing the DOIs. Technological constraints to reusing some identifiers currently exist, but we are working incrementally towards their resolution.

Sharing of identifiers between Pensoft and Plazi is currently under development and would ameliorate these problems in the future release of OpenBiodiv. While Pensoft and Plazi aim to reuse identifiers, the prefixes (plazi.org and openbiodiv.net) may remain different as both Pensoft and Plazi offer semantic information about the same objects, which may be slightly different due to the difference in focus between Plazi and Pensoft. Improved cross-linking with other platforms, like ORCID and CrossRef, would also be beneficial for uniquely identifying people and publication metadata. In addition, modifications to the XML processing workflow could help to reuse figure identifiers, since they are generated from the article DOI. Reusing figure DOIs would allow the establishment of a link between a figure and its Zenodo record via the DOI. Cross-linking would also be improved by extraction and reuse of all identifiers within Plazi's treatments.

4.3.2. User Interaction with The Knowledge Graph

Searching the OpenBiodiv database is currently possible through the SPARQL endpoint [41] or through the frontend Web portal [43]. Writing and successfully executing SPARQL queries through the endpoint can be a complex task, even with a thorough understanding of the underlying ontology structure of the graph. The Web portal addresses this difficulty by offering a frontend to the SPARQL

endpoint, which currently supports searching by a single keyword or phrase (e.g., "*Harmonia manillana*" (species) or "Agosti" (author name)). This way of searching does not allow the answering of complex questions like "How many articles about *Eupolybothrus* were published every year between 2000 and 2019?" (Table 3). A SPARQL query to answer this question is available in Table A1. Even though the Results page aims to answer more than one question about a certain entity, there are several queries that require the development of specific applications, which in turn are limited in number as the competency questions can be very diverse. Currently, the website allows modification of the SPARQL queries used for the generation of the results via the user interface, which can be utilized for educational and training purposes. In the future, different use cases and competency questions would be examined, and tools extending the search functionality of the Web portal would be created to enable an easier user interaction with the graph.

4.3.3. Enriching the Knowledge Graph

The mechanisms for information extraction from articles and taxonomic treatments currently allow the retrieval of quite an extended set of entity types (Table 1). Some extracted entities are more granular than others; for instance, the Treatment section contains subsections such as Nomenclature section and Materials Examined (Table 1). Still, there are several entity types from research articles that remain untagged or not extracted due to the complexity of modelling of such diverse information, and hence not present in the OpenBiodiv dataset. Locations, gene names, unique identifiers for collections or specimen data or specimen identifiers, especially those of type materials, are some of the most prominent examples. The extraction of these types of entities from text would significantly enrich the knowledge graph and would enable the answering of even more questions, such as "Which species, collected in location X, have been mentioned in the same or another publication?" The answer to this question could reveal potentially related organisms with various levels of biotic interactions. Markup or natural language processing of geographic, genomic and collection data would be enabled in the future by the diverse and flexible tagsets of TaxPub [35] and TaxonX [44]. This will allow even more entity types to be added to the knowledge graph.

4.3.4. Disambiguation

Information extraction is often problematic because of the ambiguity of entities. Albeit published as structured text, research articles are written in natural language, in which spelling contributes to both synonymy and homonymy. This is especially problematic for naming of people, locations, institutions and organisms. Two different spellings of the same author name (e.g., "John Smith" and "J. Smith") in an article would produce two separate entities in OpenBiodiv, each with their own unique identifier. This ambiguity would contribute to inaccurate results when a user searches for articles written by the author with multiple name spellings within the OpenBiodiv knowledge base. A similar problem exists with affiliations, which are stored as literal values, rather than identifiers (Figure 3A).

Resolving these ambiguities would require examination of related entities and the establishment of rules for uniqueness, customized for each ambiguous entity type. An increased and mandatory use of unique identifiers, for example ORCID for people [27], will make the disambiguation process much easier and straightforward.

5. Conclusions

The following conclusions can be drawn from our experience with OpenBiodiv's design and implementation.

1. OpenBiodiv is at a beta version stage but already provides a working solution to the overarching goal for creating an Open Biodiversity Knowledge Management System based on FAIR Linked Open Data.

2. OpenBiodiv serves to liberate and re-use data closed in isolated silos of biodiversity literature, including such available only in PDF.
3. OpenBiodiv allows integration of interoperable data from various sources in the biodiversity domain and federation with Linked Open Data from other domains.
4. By using open ontology and open source code, the OpenBiodiv-O is expected to catalyze Open Science principles and practices proclaimed in the Bouchout Declaration for biodiversity data.

The application of semantic technologies in OpenBiodiv helped to bridge the gap between biodiversity data and published narrative, contributing to a successful ontology modelling of the biodiversity publishing domain and resulting in the creation of the OpenBiodiv Linked Open Dataset. The enhancement of the LOD with more data types, along with disambiguation and improvements to the user interface, would strive to establish OpenBiodiv as the default Open Biodiversity Knowledge Management System.

Author Contributions: L.P. developed the idea for the project and the project plan, obtained funding, supervised the work on the project and contributed to co-writing and co-editing. M.D. contributed to co-writing and co-editing. V.S. developed the project plan, ontology, software packages and contributed to co-writing and co-editing, G.Z. was involved with the technical implementation of the project, wrote the PHP script and created the frontend website. T.G. supervised the technical implementation of the project. P.S. contributed to the ontology, project supervision and co-editing. K.S. was involved with project supervision and development of the ontology.

Funding: This research received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreements BIG4 (No 642241) and IGNITE (No 764840).

Acknowledgments: We are grateful to Plazi for contributing ideas to the project and are looking forward to many more years of fruitful collaboration.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A Example SPARQL Query

Table A1. A SPARQL query to retrieve the number of articles about Eupolybothrus which have been published every year between 2000 and 2019, along with their titles. The query can be executed at the OpenBiodiv SPARQL endpoint, where the default repository is depl2018-lite.

```
PREFIX fabio: <http://purl.org/spar/fabio/>
PREFIX prism: <http://prismstandard.org/namespaces/basic/2.0/>
PREFIX po: <http://www.essepuntato.it/2008/12/pattern#>
PREFIX openbiodiv: <http://openbiodiv.net/>
PREFIX pkm: <http://proton.semanticweb.org/protonkm#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT (COUNT (DISTINCT ?article) AS ?article_number) ?years (GROUP_CONCAT(DISTINCT
?title;SEPARATOR="; ") AS ?titles) WHERE {
    ?article a fabio:JournalArticle.
    ?article prism:publicationDate ?date.
    BIND(REPLACE(STR($date),"(\\d+)-\\d*-\\d*", "$1") AS ?year)
    FILTER (?year > "2000" && ?year < "2019")
    ?article po:contains ?tnu.
    ?article dc:title ?title.
    ?tnu a openbiodiv:TaxonomicNameUsage.
    ?tnu pkm:mentions ?scName.
    ?scName rdfs:label "Eupolybothrus"^^xsd:string.
} GROUP BY (xsd:integer(?year) AS ?years)
ORDER BY ?years
```

References

1. Agosti, D.; Egloff, W. Taxonomic information exchange and copyright: The Plazi approach. *BMC Res. Notes* **2009**, *2*, 53. [CrossRef]
2. Sarkar, I.N. Biodiversity informatics: Organizing and linking information across the spectrum of life. *Brief. Bioinform.* **2007**, *8*, 347–357. [CrossRef]
3. Hobern, D.; Baptiste, B.; Copas, K.; Guralnick, R.; Hahn, A.; van Huis, E.; Kim, E.-S.; McGeoch, M.; Naicker, I.; Navarro, L.; et al. Connecting data and expertise: A new alliance for biodiversity knowledge. *Biodivers. Data J.* **2019**, *7*, e33679. [CrossRef]
4. TDWG: History. Available online: http://old.tdwg.org/about-tdwg/history/ (accessed on 19 February 2019).
5. What Is GBIF. Available online: https://www.gbif.org/what-is-gbif (accessed on 9 May 2019).
6. pro-iBiosphere Consortium. *pro-iBiosphere—Project Final Report*; Naturalis: Leiden, The Netherlands, 2014; Available online: http://wiki.pro-ibiosphere.eu/w/media/4/46/Pro_iBiosphere_final_report_VFF_05_11_2014.pdf (accessed on 9 May 2019).
7. Senderov, V.; Penev, L. The Open Biodiversity Knowledge Management System in Scholarly Publishing. *Res. Ideas Outcomes* **2016**, *2*, e7757. [CrossRef]
8. Bouchout Declaration. Available online: http://www.bouchoutdeclaration.org/declaration/ (accessed on 9 May 2019).
9. Egloff, W.; Agosti, D.; Kishor, P.; Patterson, D.; Miller, J.A. Copyright and the Use of Images as Biodiversity Data. *Res. Ideas Outcomes* **2017**, *3*, e12502. [CrossRef]
10. Egloff, W.; Patterson, D.; Agosti, D.; Hagedorn, G. Open exchange of scientific knowledge and European copyright: The case of biodiversity information. *ZooKeys* **2014**, *414*, 109–135. [CrossRef]
11. Guralnick, R.P.; Cellinese, N.; Deck, J.; Pyle, R.L.; Kunze, J.; Penev, L.; Walls, R.; Hagedorn, G.; Agosti, D.; Wieczorek, J.; et al. Community Next Steps for Making Globally Unique Identifiers Work for Biocollections Data. *ZooKeys* **2015**, *494*, 133–154. [CrossRef]
12. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef]
13. Miller, J.; Dikow, T.; Agosti, D.; Sautter, G.; Catapano, T.; Penev, L.; Zhang, Z.-Q.; Pentcheff, D.; Pyle, R.; Blum, S.; et al. From taxonomic literature to cybertaxonomic content. *BMC Biol.* **2012**, *10*, 87. [CrossRef]
14. Page, R.D.M. Biodiversity informatics: The challenge of linking data and the role of shared identifiers. *Brief. Bioinform.* **2008**, *9*, 345–354. [CrossRef]
15. Peterson, A.T.; Knapp, S.; Guralnick, R.; Soberón, J.; Holder, M.T. The big questions for biodiversity informatics. *Syst. Biodivers.* **2010**, *8*, 159–168. [CrossRef]
16. Remsen, D. The use and limits of scientific names in biological informatics. *ZooKeys* **2016**, *550*, 207–223. [CrossRef] [PubMed]
17. Patterson, D.J.; Cooper, J.; Kirk, P.M.; Pyle, R.L.; Remsen, D.P. Names are key to the big new biology. *Trends Ecol. Evol.* **2010**, *25*, 686–691. [CrossRef]
18. Lassila, O.; Swick, R.R. Resource Description Framework (RDF) Model and Syntax Specification—W3C Recommendation 22 February 1999. W3C. 1999. Available online: https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/ (accessed on 9 May 2019).
19. Vrandečić, D.; Krötzsch, M. Wikidata: A free collaborative knowledgebase. *Commun. ACM* **2014**, *57*, 78–85. [CrossRef]
20. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web*; Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J.M., Mattern, F., Mitchell, J.C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., et al., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; Volume 4825, pp. 722–735. ISBN 978-3-540-76297-3.
21. Bingham, H.; Weatherdon, L.; Despot-Belmonte, K.; Wetzel, F.; Martin, C. The Biodiversity Informatics Landscape: Elements, Connections and Opportunities. *Res. Ideas Outcomes* **2017**, *3*, e14059. [CrossRef]
22. International Nucleotide Sequence Database Collaboration | INSDC. Available online: http://www.insdc.org/ (accessed on 11 March 2019).
23. Ratnasingham, S.; Hebert, P.D.N. Bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol. Ecol. Notes* **2007**, *7*, 355–364. [CrossRef]

24. Lepage, D.; Vaidya, G.; Guralnick, R. Avibase—A database system for managing and organizing taxonomic concepts. *ZooKeys* **2014**, *420*, 117–135. [CrossRef] [PubMed]
25. The Diptera Site. Available online: http://diptera.myspecies.info/ (accessed on 19 February 2019).
26. Hobern, D.; Apostolico, A.; Arnaud, E.; Bello, J.C.; Canhos, D.; Dubois, G.; Field, D.; Alonso García, E.; Hardisty, A.; Harrison, J.; et al. *Global Biodiversity Informatics Outlook: Delivering Biodiversity Knowledge in the Information Age*; Global Biodiversity Information Facility: Copenhagen, Denmark, 2012. [CrossRef]
27. Page, R.D.M. Ozymandias: A biodiversity knowledge graph. *PeerJ* **2019**, *7*, e6739. [CrossRef]
28. Senderov, V.; Georgiev, T.; Agosti, D.; Catapano, T.; Sautter, G.; Tuama, É.Ó.; Franz, N.; Simov, K.; Stoev, P.; Penev, L. OpenBiodiv: An Implementation of a Semantic System Running on top of the Biodiversity Knowledge Graph. *Biodivers. Inf. Sci. Stand.* **2017**, *1*, e20084. [CrossRef]
29. Senderov, V.; Simov, K.; Franz, N.; Stoev, P.; Catapano, T.; Agosti, D.; Sautter, G.; Morris, R.A.; Penev, L. OpenBiodiv-O: Ontology of the OpenBiodiv knowledge management system. *J. Biomed. Semant.* **2018**, *9*, 5. [CrossRef]
30. Wieczorek, J.; Bloom, D.; Guralnick, R.; Blum, S.; Döring, M.; Giovanni, R.; Robertson, T.; Vieglais, D. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE* **2012**, *7*, e29715. [CrossRef]
31. Peroni, S. The semantic publishing and referencing ontologies. In *Semantic Web Technologies and Legal Scholarly Publishing*; Springer: New York, NY, USA, 2014; Volume 15, pp. 121–193.
32. Catapano, T. TaxPub: An Extension of the NLM/NCBI Journal Publishing DTD for Taxonomic Descriptions. In *Journal Article Tag Suite Conference (JATS-Con) Proceedings 2010*; National Center for Biotechnology Information (US): Bethesda, MD, USA, 2010. Available online: https://www.ncbi.nlm.nih.gov/books/NBK47081/ (accessed on 19 February 2019).
33. Penev, L.; Georgiev, T.; Geshev, P.; Demirov, S.; Senderov, V.; Kuzmova, I.; Kostadinova, I.; Peneva, S.; Stoev, P. ARPHA-BioDiv: A toolbox for scholarly publication and dissemination of biodiversity data based on the ARPHA Publishing Platform. *Res. Ideas Outcomes* **2017**, *3*, e13088. [CrossRef]
34. Penev, L.; Agosti, D.; Georgiev, T.; Catapano, T.; Miller, J.; Blagoderov, V.; Roberts, D.; Smith, V.; Brake, I.; Ryrcroft, S.; et al. Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. *ZooKeys* **2010**, *50*, 1–16. [CrossRef]
35. Penev, L.; Catapano, T.; Agosti, D.; Georgiev, T.; Sautter, G.; Stoev, P. Implementation of TaxPub, an NLM DTD extension for domain-specific markup in taxonomy, from the experience of a biodiversity publisher. In *Journal Article Tag Suite Conference (JATS-Con) Proceedings 2012*; National Center for Biotechnology Information (US): Bethesda, MD, USA, 2012. Available online: https://www.ncbi.nlm.nih.gov/books/NBK100351/ (accessed on 20 February 2019).
36. Penev, L.; Lyal, C.H.C.; Weitzman, A.; Morse, D.; King, D.; Sautter, G.; Georgiev, T.A.; Morris, R.A.; Catapano, T.; Agosti, D. XML schemas and mark-up practices of taxonomic literature. *ZooKeys* **2011**, *150*, 89–116. [CrossRef]
37. RDF4R: R Library for Working with RDF. Available online: https://github.com/pensoft/rdf4r (accessed on 9 May 2019).
38. ropenbio. Available online: https://github.com/pensoft/ropenbio (accessed on 9 May 2019).
39. GBIF Secretariat. GBIF Backbone Taxonomy. Checklist Dataset. 2017. Available online: https://doi.org/10.15468/39omei (accessed on 9 May 2019).
40. OpenBiodiv. Available online: https://github.com/pensoft/OpenBiodiv (accessed on 9 May 2019).
41. Ontotext GraphDB 8.8. Available online: http://graphdb.ontotext.com/ (accessed on 15 February 2019).
42. GraphDB Workbench. Available online: http://graph.openbiodiv.net/ (accessed on 19 February 2019).
43. OpenBiodiv—The Open Biodiversity Knowledge Management System. Available online: http://openbiodiv.net/ (accessed on 19 February 2019).
44. TaxonX. Available online: https://sourceforge.net/projects/taxonx/ (accessed on 15 February 2019).
45. Pensoft Publishers. Plazi Automated Biodiversity Data Mining Workflow (Image). Available online: https://media.eurekalert.org/multimedia_prod/pub/web/164542_web.jpg (accessed on 19 February 2019).
46. Janowicz, K.; Hitzler, P.; Adams, B.; Kolas, D.; Vardeman, C. Five stars of Linked Data vocabulary use. *Semant. Web* **2014**, *5*, 173–176.

47. Bénichou, L.; Gérard, I.; Laureys, É.; Price, M. Consortium of European Taxonomic Facilities (CETAF) best practices in electronic publishing in taxonomy. *Eur. J. Taxon.* **2018**, *475*, 1–37. [CrossRef]
48. Authors Guidelines. Available online: https://zookeys.pensoft.net/about#AuthorsGuidelines (accessed on 9 May 2019).