REVIEW

by

DSc. Georgi Georgiev Markov, Professor at the Institute of Biodiversity and Ecosystem Research, BAS

of doctoral dissertation (Ph. D. thesis)

in the scientific field of *Natural Sciences, Mathematics, and Computer Science*

professional subfield *4.6 Informatics and Computer science*

Ph.D. program *Computer Science 01.01.12*

Author: *Viktor Ernestov Senderov*

Thesis title: *“The Open Biodiversity Knowledge Management System in Scholarly Publishing”*

Academic adviser: *Prof. Lyubomir Penev (IBER-BAS)*

Academic consultant: *Assoc. Prof. Kiril Simov (IICT-BAS)*

1. Description of the supplied materials

Order No. 86 from 30 .04. 2019 of the Director of IICT-BAS selected me as a member of the academic review board tasked with initiating the procedure for the defense of the Ph. D. thesis “The Open Biodiversity Knowledge Management System in Scholarly Publishing” for the academic degree ‘doctor’ in the field of “Natural Sciences, Mathematics, and Computer Science,” professional subfield “4.6 Informatics and Computer science,” Ph. D. program „Computer Science 01.01.12.” Author of the Ph. D. thesis is Viktor Senderov, a doctoral student on an independent study program within IICT-BAS with scientific adviser Prof. Lyubomir Penev and academic consultant Assoc. Prof. Kiril Simov.

Viktor Senderov has supplied the following materials:

1. Thesis (book.pdf)
2. Summary of the thesis (summary.pdf)
3. Scientific articles published as part of the dissertation

Notes: 2 papers were supplied later (RIO_article_10445.pdf, BDJ_article_10356.pdf).

4. CV

2. Short biographical data about the Ph. D. candidate

Viktor Senderov is a Marie-Skłodovska Curie Ph. D. fellow (ITN BIG4). Project beneficiary is Pensoft; education is conducted at IICT-BAS (independent Ph. D. study). His academic adviser is Prof. Lyubomir Penev (Pensoft/IBER-BAS) and academic consultant Assoc. Prof. Kiril Simov (IICT-BAS). The Ph. D. candidate has a M. Sci. degree in biostatistics from University of Munich, Germany (LMU) and a B. Sci. degree in computer mathematics from University of Magdeburg, Germany (OvGU). The thesis has been completed in an industrial setting (Pensoft) with help from the institutes IICT and IBER of BAS.

3. Novelty and adequacy of the goal and tasks

The 1992 U.N. conference Earth Summit defines biodiversity as the "diversity of living organisms of all forms including terrestrial, marine and other aquatic ecosystems, and the ecological complexes to which they belong: this includes diversity within a species, between species and within ecosystems." Studying biodiversity creates data that help shape policy both on an international and on a national scale based on knowledge for nature conservation and contributing to the advancement of the so-called "green strategy." Biodiversity of the flora and fauna of every country has considerable value as a biological resource important for the economy and agriculture. Sustainable management of these resources is important for growth in public wealth.

However, biodiversity is threatened all around the globe. Plants and animals are going extinct due to anthropogenic factors. The irreversibly of the loss of biodiversity weights heavy on the need for biodiversity research and preservation. Put into this light, the thesis topic, which includes the creation of a theoretical model (ontology) for the representation of biodiversity knowledge in machine-readable form, and besides that the creation of Linked Open Data (LOD) about biodiversity with the help of the model, is a highly relevant effort. The relevancy of the research topic is further conditioned on the fact that currently biodiversity data are spread in disparate articles and in non-interconnected databases. The importance of the topic is also connected to the fact that even though the rate of biodiversity loss is significant, which has potentially catastrophic consequences for the ecology and agriculture of Earth, there exists a taxonomic impediment: lack of funds and scientists studying biodiversity, as well as a lack of public awareness about it. From the view point of computer science, even though there are already vocabularies covering the main terms in taxonomy, the problem for inter-connecting biodiversity databases has not been solved. The same is true for the problem of knowledge extraction from biodiversity scientific papers.

4. Understanding the problem

In Introduction, Viktor Senderov introduces the scientific problem. Significant space is given to a historical overview. It should be noted that even though the standardization effort in biodiversity informatics has been going on several decades, it is still in its infancy. After the historical overview, a literature section follows. Two topics have been investigated: (1) Linked Open Data and (2) Publishing of knowledge and data about biodiversity. It is interesting to note that the candidate has reviewed very heterogeneous literature (from taxonomy, database theory, and artificial intelligence). He has approached the review process with scientific creativity and concluded: (1) Biodiversity informatics deals with highly heterogeneous data, (2) there is a lack of a system for universal identification (3) there are a lot of primary sources of biodiversity data (databases, digital versions of scientific papers) that can serve as the core of a knowledge database.

The goal and objectives of the research have been written based on these three conclusions. The goal is specified as the creation of a knowledge-management system for biodiversity based on published scientific research. This goal is subdivided into several objectives, each of which gets a dedicated chapter.

I believe that the Ph. D. candidate has reviewed the literature in depth and adequately identified a research topic.

5. Methodology

The methodology described in Introduction is modern and fits the purpose.

6. Evaluation of the thesis

Chapter 1 introduces the software architecture of the system and the term knowledge base is defined. The candidate's understanding of a knowledge base is "a database stored under a logic model that allows inference of additional facts." The main data sources are defined: digitized scientific articles published by Pensoft and Plazi, integrated with the GBIF taxonomic backbone. The selected sources cover a large amount of published biodiversity knowledge. It is very helpful that these sources already supply semi-structured data that can serve as the core of a new knowledge base.

Chapter 2 introduces the main theoretical result of the dissertation effort: the ontology OpenBiodiv-O. The chapter's text has been published (paper 8) in Journal of Biomedical Semantics.

Noteworthy is the innovative treatment of 'taxonomic concept' as a separate entity to that of 'scientific name': namely a taxonomic concept is a scientific name followed by a reference to its treatment in the literature.

Chapter 3 describes the Linked Open Data OpenBiodiv-LOD and concludes with a discussion of the Principles of Linked Open Data and the computational efficiency of the transformation process.

Chapter 4 describes the technical details and the implementation of the algorithms used to create the LOD. The implementation is based on an open source R library.

Chapter 5 describes several work-flows that deal with the import of data into the system.

Chapter 6 describes the user interface of the system. It is a web site with three target groups: simple users, specialists, and programmers. It is accessible under http://openbiodiv.net. It does not overload the user with heavy GUI details: it simply offers a search bar followed by icons for various apps. A simple search, for example, for "Lyubomir Penev" lands us on a page describing semantic information about the person. It is commendable the system accommodates different levels of users, but unfortunately several of the apps are still under development.

The remaining chapters contain source code listings, sociological data, the CV of the author and references.

7. Contributions of the thesis

In the Conclusion, several key contributions are outlined:

1. The creation of the OpenBiodiv-O ontology for publishing biodiversity data

2. The creation of the OpenBiodiv-LOD linked open data.

3. Implementation of software components needed to create the LOD.

Noteworthy are the broad spectrum of problems tackled by the thesis as well as the fact that OpenBiodiv-O bridges the knowledge representation gap between ontologies for academic publishing and ontologies for modeling of biodiversity data. The applied character of the thesis is conditioned on the development of the thesis in the industry (under an academic publisher, under the supervision of BAS). Commendable is that the research outputs are already incorporated in the business workflows of Pensoft: http://openbiodiv.net and http://graph.openbiodiv.net.

8. Evaluation of papers

The assessment was performed in accordance with the IICT-BAS Specific Requirements for calculating points for "Scientific publications in journals that are referenced and indexed in world-renowned scientific data databases (Web of Science and Scopus, Zentralblatt, MathSciNet, ACM Digital Library, IEEE Xplore and AIS eLibrary): 50 for published in Q1, 40 for posting in Q2. 30 for published in Q3, 24 for publication in Q4, 20 for published in SJR without IF and 12 for other indicators in professional field 4.6. Informatics and Computer Sciences ".

No	Authors	Published in	Index	Score
1	2	RIO/ Pensoft		0
2	21	RIO/ Pensoft	others	12
3	5	BDJ/ Pensoft	Q3	30
4	3	RIO/ Pensoft	others	12
5	13	RIO/ Pensoft	others	12
6	9	RIO/ Pensoft	others	12
7	5	ZooKeys/ Pensoft	Q2	40
8	9	J. Biomed. Semantics Springer Nature	Q1	50
9	1	Cybernetics and Inform. Tech./ BAS		0

- 3 papers (No. 3, 7, 8), published in journals indexed by SCOPUS, with SJR, and in Web of Science. 2 papers (No. 7 и 8) have impact-factor and land in Q3. Publication No. 3 and 8 are indexed in PubMed.

- 5 papers (No. 1, 2, 4, 5, 6) in RIO Journal, electronic journal publishing non-traditional research outputs. Publication 1 is not a separate scientific study but rather a Ph. D. project plan and not included in the scoring.

- One paper (No. 9) is still under review and has not been included in the scoring. It has the candidate as sole author.

All publications except 9 are published together with international collaborators and the adviser and the consultant. 4 publications have the candidate as first author (No. 1, 4, 8, 9).

23 citations have been identified, which is commendable given the short duration of the study.

9. Personal involvement of the candidate

The Ph. D. candidate has shown great independence and initiative in establishing scientific collaborations with BAS, although he is based at Pensoft. He has spent equal amount of his time on theoretical contributions, their application, and the dissemination of his scientific results as evidenced by the large number of publications. He has successfully conducted research in an interdisciplinary field with international collaborators both within and outside the Marie Curie ITN. Despite the collaborative nature of the research effort, I believe that candidate has greatly contributed to its successful completion.

10. Summary in Bulgarian

The summary in Bulgarian successfully focus on the key contributions; however, some stylistic improvements may be needed, as well as a shortening.

11. Critical recommendations

The interdisciplinary nature of the thesis conditions its broad profile and its usefulness for a broad scientific audience. I recommend to the candidate to promote the OpenBiodiv system within the biological community.

12. Personal impressions

My personal contact with Viktor Senderov has left a positive impression. I know him from an exam and form several of his presentations.

13. Recommendation for further use of the results

It is crucial that the website and dataset remain publicly accessible. In a greater philosophical sense, biodiversity can be regarded as knowledge accumulated through the evolution of species in the course of millions of years about how to survive on Earth. It is especially troublesome to consider that this "Library of Knowledge" is presently being burned down. Taxonomic, trophic, functional, genetic and other dimensions of biodiversity are still relatively unknown. Even taxonomic knowledge, the most investigated aspect of biodiversity is not complete and heavily biased towards the species level, megafauna, and organisms relevant for humans. Helping to organize the knowledge with the OpenBiodiv system will facilitate the rational description of biodiversity on Earth

CONCLUSION

The Ph. D. thesis contains scientific and applied results that constitute an innovation and fulfill the requirements of "Rules for the Implementation of the Law on the Development of the Academic Staff in the Republic of Bulgaria" and its "Regulations for Implementation". The supplied materials comply to the "The Regulations for the Special Conditions for Acquisition of Academic Degrees and the Possession of Academic Positions in the Institute of Information Technology and Communications, BAS". The thesis shows that the candidate has in-depth theoretical and professional knowledge in the field of Computer Science and has shown capability of independent research.

Given the stated facts, I convincingly give my positive review of the thesis, as evidenced by the materials (book, summary, papers) and advise the jury to award the degree 'doctor' to Viktor Senderov in the field "Natural Sciences, Mathematics and Informatics", subfield "4.6 Informatics and Computer Science", doctoral program "Computer Science 01.01.12"

29. 05. 2019

Reviewer: Prof. DSc. Georgi Markov