REVIEW

of dissertation work for covering of the educational and scientific degree PhD in the field of higher education 4. "Natural Sciences, Mathematics and Informatics", Professional Field: 4.6 "Informatics and Computer Science", Scientific specialty: 01.01.12 "Informatics"

PhD student: *Victor Senderov*

Title: *The Open Biodiversity Knowledge Management System in Scholarly Publishing*

Supervisors:
Prof. Lyubomir Penev and Prof. Kiril Simov

Reviewer: *Assoc. prof. Svetla Boytcheva – IICT-BAS*

This review was written and presented on the basis of Order 86 from 30.04.2019 of the Director of IICT-BAS in compliance with the decision of the Scientific Council of IICT-BAS (rec. of proceedings No. 5 from 24.04.2019), as well as the decision of the Scientific Jury under the Procedure (rec. of proceedings No. 1 from 10.05.2019). It was written on the basis of the Act of the Development of the Academic Personnel in Republic of Bulgaria (ADAPRB), Act No. 26 from 13.02.2019 for the amendment and supplementation of the Rules of implementation of ADAPRB (RIADAPRB), concerning the field 4. Natural sciences, mathematics and informatics, Professional Fields 4.1., 4.2. 4.3., 4.4., 4.5., 4.6., Rules on the conditions and order for acquiring academic degrees and occupying academic positions at the Bulgarian Academy of Sciences, Regulations on the Specific Conditions for Acquisition of Academic Degrees and for Applying Academic Staff in the IICT-BAS.

1. Comprehensive analysis of the scientific and scientific-applied achievements in the dissertation work. Characterization of key achievements.

The dissertation is written in English and has a volume of 113 pages, including 15 pages of references, 2 pages of author's reference, 4 pages of listings, 1 appendix of 2 pages, and 2 blank pages. Due to the specificity of the selected format, in view of the fact that the dissertation will be printed by Pensoft, it should be noted that there are deviations from the typing page standard (38 lines, 66 symbols) - 51 lines, 85 symbols, the standard equals about 150 standard pages, without counting the references list, author's reference, and appendices. The text is organized in introduction and 8 chapters and conclusion. The references list consists of 164 publications, including 9 publications of PhD student. The majority of cited literature is in English, and there are several sources in Russian, French, and Latin. Most of the cited publications have been published over the last 10 years. The references are appropriately cited in the dissertation text.

Actuality of the problem

The dissertation topic is related to a very up-to-date and fast growing topic - creation and management of linked open data (LOD - https://lod-cloud.net/). The presented scientific work

describes the design and creation of an OpenBiodiv system for working with LOD on biodiversity. This topic is very important because there is a strong need to systemized knowledge in the field, both as relationships between the different concepts organized in taxonomies and ontologies, as well as terms names alignment and agreement, and to create a digital repository of open linked data.

Related Work

The presented analyses and overview of knowledge based databases and related open data, as well as the presented data survey in biodiversity, show that the PhD student has deep and systematic knowledge about the problem. Based on the presented state of the art study, the OpenBiodiv project was defined and initiated. The related work presented in the PhD thesis also show that the PhD student knows the domain and the main research problems that can address in it.

Goal and objectives

The aim of the dissertation is to create an Open Biodiversity Knowledge Management System (OBKMS), focusing on knowledge extracted from scientific literature.
To achieve the objectives of the dissertation, 6 tasks have been formulated: (1) ontology creation; (2) design of system architecture, (3) creation of linked open data (LOD) on the basis of published taxonomic articles using the ontology defined in (1), (4) Develop methods for converting taxonomic publications into the semantic model of the ontology in order to support (3); (5) Develop practical workflows for continuously converting taxonomic data into taxonomic publications and thus updating the LOD dataset, (6) Create a web portal and example applications on top of the knowledge base.

The primary model for OBKMS is based on data representation as a semantic knowledge graph, and GraphDB (https://www.ontotext.com/products/graphdb/).

Analysis of the achievements presented in the dissertation work

The introduction section contains an overview of the subject area, major research problems identification and definition of the goal of the dissertation, within 6 main tasks. The dissertation thesis describes the OpenBiodiv project, for the Open Biodiversity Knowledge Management System (OBKMS) through the creation of an open knowledge-based biodiversity information system derived from the scientific literature. A justification for the chosen research methodology was made. Different aspects are considered:

- semantic graphs and RDF schema presentation are selected for knowledge presentation;
- the choice of information sources - the Pensoft publication, which is based on publications with enhanced semantic information (enhanced publications); GBIF Backbone Taxonomy; and Plazi Treatment Bank.
- A flexible software development method (Agile) is selected.

The structure of the dissertation is also presented.

Chapter 1 describes the architecture of OpenBiodiv - (1) semantic graph database - ontologies, open linked data; (2) back-end - methods for converting unstructured information into a form that allows its semantic links to be represented, (3) a front-end part - a web portal. A flexible software

development method (Agile) has been used to design the system. The implementation is based on GraphDB.

Chapter 2 presents the OpenBiodiv-O ontology. The domain conceptualization process is explored and described in detail, so that this process can be automated in the creation of the OpenBiodiv-O ontology. An overview has been made of the related work in this area. A semantic model of the Biodiversity Publishing Domain is presented. The defined classes and relationships between objects in these classes are represented. The semantic modeling of the biological nomenclature is also presented. There is also described a hybrid hierarchy of classes that accommodate both traditional taxonomic name usages and the usage of taxonomic concept labels and operational taxonomic units

Chapter 3 describes OpenBiodiv-LOD (Open Linked Data) generation by processing three major sources of information: GBIF Backbone Taxonomy; postings from Pensoft; and Plazi Treatment Bank. OpenBiodiv-LOD is a synthetic set of data and does not contain any new data but only integrates data contained in the information sources. A pseudocode of the Transformation Extractor Algorithm from XML to RDF is also provided.

Chapter 4 describes the RDF4R library, which is R library for working with RDF. Described are the main features of the package. An example is given for converting a SPARQL query into a function of R.

Chapter 5 discusses the automated workflow for processing data flow related to biodiversity. Two methods are presented: (1) Automated specimen record import, and (2) automatic generation of ecological metadata language (EML).

Chapter 6 describes the web portal http://openbiodiv.net/ . The functionalities of the system can be used by the three types of users: (1) basic level: using the workspace with tools, (2) expert level: using applications and (3) expert: using SPARQL or R. Still not all search tools have been developed. The web portal offers the basic semantic search functionality for the three types of users.

Chapter 7 contains source code listing of some SPARQL queries that are described in the previous chapters. In addition, the source code of the RDF4R library has been set.

Chapter 8 describes the presentation of the iDigBio Webinar. The listener profile is described.

The **conclusion** contains an assessment of the fulfillment of each of the assigned tasks in the dissertation. Suggested guidelines for future work are outlined. An author's achievments, publications, citations, approbation of the results - exported reports are attached. Altogether 8 publications were submitted, 20 citations were noted. One of the publications is a work plan for the dissertation. Results of the dissertation were presented at 3 internal scientific seminars at the Bulgarian Academy of Sciences and 16 international events.

The contributions

There are 3 main contributions of the PhD student - one scientific and two in the area of applied sciences:

Scientific contribution: • Creation of OpenBiodiv-O ontology and a formal model of the field of biodiversity knowledge publishing.

Applied sciences contributions:

• Related work overview and analyses and development of OpenBiodiv-LOD

• Implementation of OpenBiodiv software modules

Reliability of the achieved results

The main focus of all presented research is the openness of the data and the program code. Links to all datasets, resources, and program implementations are provided. The eight publications of the PhD student, the overview of the related work on the dissertation theme, the direct participation of the doctoral student in the research process and the description of the studies made are also covered by the dissertation.

2. Publications

From a formal point of view, in accordance with the Regulations for the implementation of the ADAPRB, Decree No 26 of 13.02.2019 for the amendment and supplement of the RIADAPRB, field 4. Natural sciences, mathematics and informatics, Prof. scientific fields 4.1., 4.2., 4.3. , 4.4., 4.5., 4.6., And the Regulations for the Special Conditions for Acquisition of Academic Degrees and Academic Degrees in IICT-BAS, the PhD student is required to have at least 30 points from criteria in Group G (indicators from 5 to 10).

PhD candidate Viktor Senderov has presented 8 articles in total, all of which are in English, open access publications, including:

- 3 publications (No. 3,7,8) in the SCOPUS reference journals, with SJR, in addition all 3 publications are also referenced in the Web of Science, two of which (No. 7,8) also have an impact factor and fall into quartile Q3. Publications (No. 3,8) are also indexed in PubMed
- 5 publications (No. 1, 2, 4, 5, 6) are published in Research Ideas and Outcomes - Pensoft's online journal. Publication No. 1 is not a scientific publication, but a work plan for the doctorate, so it will not be included in the calculation of the points.

All presented publications are co-authored both by Viktor Senderov's supervisors, as well by Bulgarian and foreign scientists. He is the first author of three publications (No. 1, 4, 8). The PhD candidate has no publication as single author.

According to the Amendment No. 26 of 13.02.2019 of RIADAPRB, only scientific publications which are referenced and indexed in international scientific information databases (Web of Science and Scopus, Zentralblatt, MathSciNet, ACM Digital Library, IEEE Xplore and AIS eLibrary) fulfill the requirements of the criteria in Group G.

Thus, only 3 of the publications (No. 3, 7, 8) can be considered for the criteria in Group G (indicators from 5 to 10). Then the total number of points for this indicator is 80 pts, which also meets the minimum requirement of 30 points.

№	Journal	Publisher	Indexing and referencing	Points
3	Biodiversity Data Journal	Pensoft	WoS SCOPUS, SJR 0.465	**30**
7	ZooKeys	Pensoft	WoS IF 1.079, Q3 SCOPUS, SJR 0.533	**30**
8	Journal of Biomedical Semantics	Springer Nature	WoS IF 1.6, Q3 SCOPUS, SJR 0.952,	**20**
			Total	**80**

3. Citations

There are 18 citations of the PhD publications presented in the dissertation. Publications in reputed international scientific journals, as well as their numerous citation by world scientists, show the importance of the results of scientific research in the field.

4. Contribution in co-authored publications

In my opinion, the candidate's contribution to collective publications is clear and substantial. It is evident from the subject matter. In three of the publications, the PhD student is the first author.

5. Critical notes

I would like to make some technical notes that do not decrease the value of the value of the presented results in the PhD thesis, but can help to improve their presentation:

1. It would be better to separate print from electronic sources in the reference list.
2. The text of the PhD thesis needs polishing of the presentation style and terminology. There are some punctuation and spelling errors, the removal of which would only help to improve the overall impression of the presented scientific results. It would be good to avoid the use of strangers and jargon words.
3. In general, the structure of the dissertation text does not correspond to the traditionally accepted layout. There is no chapter overview - there are only 3 pages of literature review in the introductory part, but they also include a description of the structure of the dissertation. Code listings (Chapter 7) should be presented as an appendix after the main dissertation text. The appendix (Chapter 8) should also not be considered as a part of the main text. There is no list of figures and a list of tables, which are usually included in the description.
4. In my opinion, the aim of dissertation is not very well defined - it needs some refinement in the direction of focus on the scientific problem that is being solved.
5. I would like to recommend to the Ph.D. student to publish individual scientific publications in the area 4.6 Informatics and Computer Science, in order to promote the results and to obtain an objective evaluation of his achievements in this field by international reviewers.

6. PhD abstract

The presented PhD abstract in Bulgarian contains 54 pages and reflects the main chapters and results of the dissertation.

7. Conclusion

In my opinion the presented publications and the results described in the PhD thesis show that the PhD candidate Victor Senderov complies the requirements of the ADAPRB, the RIADAPRB, and the specific requirements of the regulations of BAS and IICT-BAS. The minimum national and IICT-BAS requirements for awarding the PhD in. 4.6 Informatics and Computer Science.

I give my positive conclusion for the award of educational and scientific degree PhD to Viktor Senderov in the field of higher education 4. "Science, Mathematics and Informatics", field 4.6 "Informatics and Computer Science", scientific specialty: 01.01.12 "Informatics".

Sofia, 7 June 2019 г.

Reviewer:

/Assoc. Prof. Svetla Boytcheva/