Research Ideas and Outcomes 2: e7757
doi: 10.3897/rio.2.e7757

OPEN ACCESS

PhD Project Plan

The Open Biodiversity Knowledge Management System in Scholarly Publishing

Viktor Senderov‡, Lyubomir Penev‡

‡ Pensoft Publishers & Bulgarian Academy of Sciences, Sofia, Bulgaria

Corresponding author: Viktor Senderov (datascience@pensoft.net)

Reviewable v1

Received: 11 Jan 2016 | Published: 11 Jan 2016

Citation: Senderov V, Penev L (2016) The Open Biodiversity Knowledge Management System in Scholarly Publishing. Research Ideas and Outcomes 2: e7757. doi: 10.3897/rio.2.e7757

Abstract

This project aims to develop and implement novel ways of publication, visualization, and dissemination of biodiversity and biodiversity-related data and thus bring the Open Biodiversity Knowledge Management System closer to fruition. In order to do so, we will develop new types of Enhanced Publications (EP's), which will allow automated data import into the manuscript and export from the manuscript and provide dynamic visualizations. These EP's will enable biodiversity researchers and taxonomists to streamline their work and publish more data-rich species descriptions.

Keywords

entomology, systematics, taxonomy, biodiversity informatics, evolutionary informatics, bioinformatics, data visualization, data publishing, semantically enhanced publication

Objectives, concept and approach

This PhD project plan constitutes a translation and an expansion of the original Bulgarian version titled "Публикуване, визуализация и разпространение на първични и геномни данни за биологичното разнообразие на основата на открита система за

управление на информацията" - "Publishing, Visualizing, and Dissemination of Primary and Genomic Biodiversity Data Based on the Open Biodiversity Knowledge Management System," which was officially approved at the Bulgarian Academy of Sciences on 27. Oct. 2015.

Up to the time of writing of this PhD project plan, willingness to create an Open Biodiversity Knowledge Management System (OBKMS) was declared by over ninety institutional and many more individual signatories of the Bouchout Declaration. The goals and purpose of the system were set forth in the project deliverables from the pro-iBiosphere project (pro-iBiosphere 2014a, pro-iBiosphere 2014b). A number of articles have been previously published on the topics of linking data and sharing identifiers (Page 2008), unifying phylogenetic knowledge (Parr et al. 2012), taxonomic names and their relation to the Semantic Web (Page 2006, Patterson et al. 2010), aggregating and tagging biodiversity research (Mindell et al. 2011) and other topics included in the scope of the OBKMS. Some discussion on the OBKMS is to be partially found in the science blog iPhylo (Page 2014, Page 2015). The legal aspects of the OBKMS have been discussed by Egloff et al. (2014). However, until now, a detailed scientific discussion of the OBKMS on its own, together with its implementation consisting of algorithms, protocols, tools, case-studies and so forth, has not been published in the peer-reviewed literature. Therefore, there is a definite need for a fuller scientific treatment of the OBKMS. This dissertation aims to treat the OBKMS in the area of scholarly publishing. In the introductory part of the dissertation we want to define what we understand OBKMS to mean (a very concise and accurate specification that can be interpreted unambiguously), and then design and implement parts of it, which relate to its publishing aspects, in the subsequent chapters.

According to pro-iBiosphere (2014a), biodiversity and biodiversity-related data have two different "life-cycles." In the past, after an observation had been made, it was recorded on paper and then it, or a summary of it, was published in paper-based form. In order for this data to be available to the modern scientist, efforts are made nowadays to digitize those paper-based publications (Agosti et al. 2007, Miller et al. 2012). For this purpose, several dedicated XML schemas have been developed (see Penev et al. 2011 for a review), of which TaxPub and TaxonX seem to be the most widely used (Catapano 2010, Penev et al. 2012). The digitization of those publications contains several steps. After their scanning and OCR, the process of text and data mining starts usually combined with a search for particular kinds of data, which 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). At present, data and publications are mostly "born digital." I.e. data is measured and recorded in a digital format and stored in a database (Smith et al. 2013). It is then utilized for research published in electronic format. The challenge here is to model and integrate the many types of data that are currently being acquired and published in the biodiversity and bioinformatics domains. Another challenge is to set the data into context. This can be achieved by modeling the relationships between the objects and exporting the data as Linked Data (Berners-Lee et al. 2001). This dissertation project will deal primarily with the life-cycle of this new emergent type of the born-digital data.

Several tools and systems that deal with the integration of biodiversity and biodiversity-related data have been developed by different groups. Some of the most important ones are UBio, Global Names project, BioGuid, BioNames, Pensoft Taxon Profile, the Plazi Treatment Repository, and others.

According to the afore-mentioned pro-iBiosphere brochure, the OBKMS must be built upon ten principles:

1. Consistent biodiversity information space.
2. Open and accessible research data.
3. Collaboration in data collection and reuse.
4. Convenient provision of high quality information.
5. New formats to support novel and diverse uses.
6. Prognostic modeling.
7. Linkages with other resources.
8. Improved markup processes and infrastructure.
9. Literature discovery and conversion.
10. Accreditation for researchers' work.

The objective of this dissertation project will therefore be to study, develop and apply new types of enhanced electronic publications that implement the principles of the OBKMS and aid in the publishing, visualization and re-use of research data and its associated narrative.

Project description

One of the main challenges of the OBKMS is to develop a system for robust and universal identification of biodiversity and biodiversity-related objects, such as taxon names, taxon name usages, museum specimens, occurrence records, taxon treatments, genomic sequences, organism traits, bibliographic citations, figures, multi-media files, etc. Historically, many such systems have been proposed and utilized. For example, Darwin Core Triplets, the de facto standard for occurrence-type data are discussed in Guralnick et al. (2014). Globally Unique Identifiers (GUID's) such as Life Science Identifiers (LSID's) are defined and discussed in Page 2008, Pereira et al. 2009, Richards 2010. Universal Resource Names (URN's), Universal Resource Identifiers (HTTP-URI's), and Digital Object Identifiers (DOI's), are all discussed in Guralnick et al. 2015 and others. However, this multiplicity of systems has led to a situation where no truly universal system has been adopted (Guralnick et al. 2015). Moreover, for systems that have been more widely adopted, such as Darwin Core Triplets for occurrence and specimen data, Guralnick et al. (2014) observed significant difficulties in cross-linking the same specimen across different databases. Therefore, one of the tasks of this work will be to propose and implement a system for the robust identification and access of biodiversity-related data in a taxonomic or a bioinformatics publication. Furthermore, we aim at creating a data model of these data objects that allows researchers to address important scientific questions.

We are of the opinion that the OBKMS 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 (Nature 2015).

One such instrument that we plan to utilize is the Enhanced Publication (EP) (Claerbout and Karrenbach 1992, van Godtsenhoven et al. 2009, Shotton 2009). According to the first source, "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."

In other words, the act of publishing in a digital, enhanced format, differs from the ground up from a paper-based publication. The main difference is that the document can be structured in such a format as to be suitable for machine processing and to the human eye. In the sphere of biodiversity science, journals such as ZooKeys, PhytoKeys, and the Biodiversity Data Journal (BDJ) in particular, have already made first steps in the direction of EP's (Penev et al. 2010).

EP's can be connected to one of the main issues facing zoology nowadays, which is the discrepancy between traditional morphologically described species and the growing number of species delimited via genomic technologies (Page 2011). On the one hand, a huge amount of described species do not have genomic (barcode) data associated with them, the so-called *known unknowns* (Collins and Cruickshank 2014); on the other hand, there is a big and growing number of species, delimited only based on genomic information and grouped as Operational Taxonomic Units (OTU's) (Ratnasingham and Hebert 2013) or Species Hypotheses (SH) (Kõljalg et al. 2013). These unnamed species are called *dark taxa (*Page 2011) or *unknown knowns* (Collins and Cruickshank 2014). If we do not manage to reconcile these two worlds, we face the danger of losing centuries of accumulation of information by generations of scientists on the morphology, biology, behavior and other characteristics of most of the currently known species.

The large number of dark taxa is due to the fact that genomic technologies are very effective and allow for the generation of SH or OTU's with a speed much higher than the speed with which taxonomists manage to publish morphological descriptions and name them (Page 2011). The solution for this problem may come if modern systems are created that enable the semi-automatic generation of taxonomic manuscripts in which authors are able to rapidly publish species descriptions and link them to OTU's. Such approaches are known as "turbo-taxonomy" (Butcher et al. 2012), i.e. the description of species with morphological and genomic data straight into an enhanced digital format. Therefore,

another task of the work would be to develop EP's enabling the rapid description and re-description of species together with information on their genomes.

EP's can also be connected to another interesting issue in bioinformatics - namely the publication, visualization and analysis of genomic data. Recently, interest in data visualization in the genomic and publishing communities has risen sharply: there have been blog posts about visualizing phylogenies (Page 2015) and new data journals, offering enhanced publications with visualizing capabilities have been launched (Brill and DANS 2015). As part of the BIG4 International Training Network, of which this dissertation project is a member of, a project will be dedicated solely to the visualization of genomic data and will be supervised by the BIG4 partner ERA7. We see, therefore, a great opportunity for collaboration with the BIG4 project partners in order to develop visualizations of genomic and other biodiversity-related data.

One example of such a visualization as part of an enhanced publication may be a very large phylogeny (Page 2015). It is obvious that phylogenies are one of the main outcomes of many papers in evolutionary biology. A quick search in Google Scholar for papers containing the word "phylogeny" and published since 2015 returns about 21,000 results. As we live in the era of big data, and as constructed phylogenies become bigger and bigger (Hinchliff et al. 2015), the feasibility of displaying them in a static image format is diminishing. We believe that the future of publishing phylogenetic articles lies in providing the scientists with interactive plots which enable them to display thousands of clades in a single diagram by way of zooming in and out, showing and hiding parts of the diagram and/or displaying graphical summaries of cladistics relationships. Therefore, one avenue of research for this PhD project in the part on visualization might be displaying interactively very big phylogenies as a part of an EP. It should be noted that many similar initiatives exist, such as the Open Tree of Life, and that efforts should be made not to duplicate them.

Another related example is the graphical display of OTU and SH data. Since genomic methods for species delimitation provide different outcomes depending on the selected cut-off value for the similarity (Kõljalg et al. 2013), methods are needed for displaying SH interactively as part of an EP, whereby the user might select the number of SH to be displayed and/or the cut-off value. Therefore, another avenue of research for this PhD project in the part on visualization might be displaying interactively OTU's and SH.

Yet another example, related to the previous two, is displaying metagenomic data. In metagenomic data, sequence information from the environment coming potentially from many different species is mixed together (Hugenholtz and Tyson 2008). By using barcode genes, the biodiversity of the sample may be assessed (Smith and Fisher 2009). A graphical summary of this information will greatly simplify the researcher's tasks. We suggest to collaborate with the aforementioned ERA7 to develop a workflow which will enable researchers to publish their metagenomic samples as an EP and then use the displays developed at ERA7 to visualize their results. More specifically, we envisage the creation of a special type of genomic or metagenomic data paper geared towards the need for rapid publication and visualization of sequence data as a scholarly article.

Finally, in order to complete the full life-cycle of the data as described in the OBKMS brochure (pro-iBiosphere 2014b), we would like not only to import and display data in the manuscript, but also to export the data with machine computability and "processability" in

mind. To accomplish this task we view the individual data records mentioned in the publication as objects with universal identifiers, and offer a web service designed for the retrieval of each of these objects. This service can then be used to map these objects and their relationships to entities within the Semantic Web (Berners-Lee et al. 2001) and to integrate these mappings as Linked Open Data into Semantic Web-compatible databases such as Wikidata (Mitraka et al. 2015). Wikidata is a good choice of an external database, as it is a well-established and sustainable platform suitable to be used by the research community for sharing data and results (Mietchen et al. 2015).

This approach is complementary to the approach taken by ContentMine. While ContentMine uses text-mining to find "facts" within thousands of articles of scientific literature, we will start by taking semantically enriched publications where pieces of data can easily be identified. The goal of both projects is to export the data to the Linked Open Data cloud.

One possible way of implementing this workflow could be the extraction of the data objects from the manuscripts and their storage in a database. In order to capture the complex relationships between the various biodiversity-related objects, a graph data model might be most appropriate (Neo4j 2015, Pareja-Tobes et al. 2015). Expertise in graph database technologies is also present in the BIG4 project partner ERA7.

Implementation

Through the implementation of the work-tasks described in this section, we would like to build and describe the implementation of a part of the OBKMS in the area of publishing of digitally born biodiversity literature.

Methodology

Materials will be collected together with other BIG4 project partners in different expeditions. Also museum and database records will be used. For the realization of the technical part the following internet technologies will be utilized:

- semantic publishing (XML, ontologies),
- web technologies (PHP, node.js) and others.

For the realization of the scientific part in bioinformatics, the following methods will be utilized:

- technologies for sequencing and analysis of genomic data including Next Generation Sequencing,
- algorithms from evolutionary genetics and phylogeny, statistical algorithms (e.g. MrBayes, cluster analysis),
- interactive visualizations (d3.js).

An array of methods is to be expected as the output of the work itself. We also intend to realize the work as an open thesis. This means that we intend to open the access to the primary scientific output (scientific papers) written as a result of the effort. Also, we will try to open a maximum amount of secondary scientific output such as lab notebooks, software code, etc. Finally, we will aim at involving the BIG4 community, as well as the wider scientific community, in contributing to the discussion by means of a popular blog, which can be found at http://openbkms.blogspot.com.

Work plan

Work-task 1: Propose, model and realize a software system for universal identification, access and handling of sub-article level data elements such as article metadata, article sections, taxon names, taxon treatments, collection specimens, occurrence records, genomic sequences, species traits, images, tables, and so on.

Sub-tasks:

- Specify and design a system for accessing, harvesting, querying, and archiving of sub-article level biodiversity-related data in an Enhanced Publication.
- Use this specification and design to implement a software system to achieve these means.

Work-task 2: Develop, test, and apply new forms of EP's, allowing for automated or semi-automated exchange of data with international biodiversity portals such as GBIF, NCBI, IUCN, ZooBank, UNITE, iDigBio, DataONE, and others.

Sub-tasks:

- Analyze existing algorithms for species delimitation in Operational Taxonomic Unit's (OTU's) or Species Hypotheses (SH), and based on them, propose a model for fast and modern formal taxonomic descriptions in an open publishing platform and in machine-readable format.
- Create a system for exchange of data between the publication platform (BDJ) and biological databases (e.g., GBIF, iDigBio, GenBank, BOLD, PlutoF, etc.).
- Create new formats for scientific publications aimed at specific and non-traditional research outputs such as:
    - data papers (for taxonomic and ecological data),
    - data papers (for genomic data),
    - species conservation profiles,
    - alien species profiles,
    - species genomic profiles.
- Test and potentially apply the idea of nano-publications via modeling the information found in article abstracts.
- Investigate the possibility of integrating source code with an EP (literate programming) in the area of biodiversity informatics. As an example, one can

investigate the integration of algorithms for forecasting and visualizing the distribution of invasive species.

Work-task 3: Develop and integrate new methods for publishing and visualizing of genomic and metagenomic data with with platforms such as BDJ.

Sub-tasks:

- Investigate algorithms for visualization of phylogenetic trees.
- Display gene information profiles for specific genes as an enhancement of a publication.
- Investigate algorithms for visualization of metagenomic data.

Work-task 4: Apply the novel methods from Work-tasks 1, 2, and 3 to publish one or more pilot publications together with BIG4 project partners.

Educational program:

Courses that will be attended, and deadlines:

- First BIG4 workshop in Copenhagen at the kick-off meeting (September 2015).
- Language and computer science courses at the Bulgarian Academy of Sciences (2015 - 2016).
- Specialized course at the Bulgarian Academy of Sciences: "Taxonomy and phylogenetics" (Oct. - Nov. 2015).
- Specialized course in taxonomy as part of the first BIG4 summer school: "Introduction to diversity and collecting of big four insect groups and one day complementary training in PhD project management for ESR's" (May 2016)

Other BIG4 courses which may be attended:

- Second workshop (Turku, 10 days): "Training course in laboratory data capture (genomics) with the participation of a visiting DNA scientist."
- Third workshop (Jena, 10 days): "Training course in laboratory data capture (morphology)."
- Second summer school (14 days, Stockholm) including two training courses: "1. Statistical phylogenetics and phylogenomics (basic) (7 days); 2. IT solutions for citizen science (7 days)."
- Fourth workshop (14 days, Stockholm) including two training courses: "1. Statistical phylogenetics and phylogenomics (advanced) (10 days); 2. Cloud computing (4 days)."
- Fifth workshop (10 days, Sofia) including two training courses: "1. Advanced biodiversity and data publishing, including copyright issues (7 days); 2. Business training on how to start a company (3 days)."
- Third summer school (10 days, Vienna) including three training courses: "1. Grant writing (ERC, Marie-Curie, national grant agencies (3 days); Future career planning (2 days); Work/life balance (2 days)."

Individual study will be done leading up to an exam at Bulgarian Academy of Sciences covering the following topics:

- Biodiversity informatics: basics of gathering, analyzing and publishing primary biodiversity data; basics of taxonomy and phylogenetics; barcoding and molecular phylogeny.
- Semantic methods and technologies for data exchange.
- Bioinformatics and genomics: methods for analyzing, visualizing and publishing of genomics data; cloud computing; methods for analysis of metagenomes.
- Promotion and PR of open science.

Details for replicability and reproducibility

Embedding of source code in the EP's will allow readers to test the results by running and modifying data and code. The thesis will be developed in accordance with the open science approach, which assumes that all data and standalone software tools or code developed through the project will be available as open data and open source.

Timeline

Work program for 2015

1. Prepare bibliographic reference and read introductory literature.
2. Attend base courses at the Bulgarian Academy of Sciences.
3. Prepare a PhD project plan.
4. Begin work on Work-tasks 1 and 2.
5. Participate in conferences:
    1. EU-BON in Cambridge, UK.
    2. EMODNET in Crete, Greece.
    3. BExIS in Jena, Germany.
    4. BIG4 kick-off meeting in Copenhagen, Denmark.
6. Create the blog.

Work program for 2016

1. Attend BIG4 field seminar in the Czech Republic (Pec pod Sněžkou).
2. Complete work on Work-tasks 1 and 2.
3. Start work on Work-task 3.
4. Take candidate exam.
5. Prepare one or two open access publications describing the results of Work-tasks 1 and 2.
6. Participate in further conferences and maintain blog.

Work program for 2017

1. Secondment in genomics at a BIG4 project partner.
2. Complete work on Work-task 3.
3. Prepare an open access publication describing the results of Work-task 3.
4. Co-author publications together with BIG4 project partners (Work-task 4).
5. Participate in conferences and maintain the blog.

Work program for 2018

1. Combine publications into a dissertation.
2. Submit dissertation.
3. Defend the dissertation.
4. Prepare further research career.

Expected results and impact

As part of the scientific and methodological results, we expect to develop new approaches, methods and formats for publishing of data and narrative in biodiversity science. We also expect to develop novel methods for information flow between publications and external data repositories, and to illustrate the aforementioned methods in exemplar papers using data gathered in the BIG4 consortium.

We also expect a minimum of two scientific publications in open access journals, where the student will be the first author:

- Paper dealing with the results of Work-task 1; preliminary title: "Free the data: publishing data as part of an Enhanced Publication in biodiversity science."
- Paper dealing with results of Work-task 2; preliminary title: "A concept and methods for data exchange between publication and data provider in the context of open science."
- Papers dealing with Work-task 3 depending on the results.

Furthermore, methods and other scientific exemplar papers are expected where the the student will be a co-author. The student will also give presentations at international symposia, write blog posts and actively popularize the results in the social media. The open science approach to development of the PhD, which starts with publication of the present PhD research plan, will be utilized throughout.

Acknowledgements

We would like to thank our colleagues at Pensoft, in particular Pavel Stoev and Teodor Georgiev - for many helpful working meetings; my colleagues at the Bulgarian Academy of Sciences - for reviewing and approving the Bulgarian version of this work plan; our partners

at Plazi, in particular Donat Agosti and Terry Catapano - for contributing ideas and useful discussions; the reviewers, Alexey Solodovnikov, Daniel Mietchen, and Donat Agosti - for insightful comments and suggestions for improvement; and Prof. Rod Page - for the stimulating discussions and advice.

Funding program

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 542241.

Project

This PhD project is being developed by Mr. Viktor Senderov under the supervision of Prof. Lyubomir Penev at Pensoft Publishers and the Bulgarian Academy of Sciences as part of the larger BIG4 (Biosystematics, Informatics and Genomics of the 4 big insect groups) EU training network. The dissertation will be defended at the Bulgarian Academy of Sciences. In case of successful defense, Mr. Viktor Senderov will be awarded the title of Doctor of Entomology, whereas the field of specialization is Bioinformatics.

Hosting institution

Pensoft Publishers, Bulgarian Academy of Sciences, various BIG4 partners.

Author contributions

Viktor Senderov participated in the concept development of his PhD and wrote the main text of the paper.

Lyubomir Penev developed, together with Viktor Senderov, the overall concept of the PhD plan, edited and revised the manuscript.

References

- Agosti D, Klingenberg C, Sautter G, Johnson N, Stephenson C, Catapano T (2007) Why not let the computer save you time by reading the taxonomic papers for you? Biologico 69: 545-548.
- Berners-Lee T, Hendler J, Lassila O (2001) The Semantic Web. Scientific American 284 (5): 34-43. DOI: 10.1038/scientificamerican0501-34

- Brill, DANS (2015) Brill and DANS Launch a New Open Access Journal on Research Data. http://www.brill.com/news/brill-and-dans-launch-new-open-access-journal-research-data. Accession date: 2015 11 13.
- Butcher BA, Smith MA, Sharkey M, Quicke DJ (2012) A turbo-taxonomic study of Thai *Aleiodes (Aleiodes)* and *Aleiodes (Arcaleiodes)* (Hymenoptera: Braconidae: Rogadinae) based largely on COI barcoded specimens, with rapid descriptions of 179 new species. Zootaxa 1: 232. URL: http://www.mapress.com/zootaxa/list/2012/3457.html
- Catapano T (2010) TaxPub: An Extension of the NLM / NCBI Journal Publishing DTD for Taxonomic Descriptions. Proceedings of the Journal Article Tag Suite Conference 2010. Proceedings of JATS Conference, 1-6 pp.
- Claerbout J, Karrenbach M (1992) SEG Technical Program Expanded Abstracts 1992. 4 pp. URL: http://dx.doi.org/10.1190/1.1822162 DOI: 10.1190/1.1822162
- Collins RA, Cruickshank RH (2014) Known Knowns, Known Unknowns, Unknown Unknowns and Unknown Knowns in DNA Barcoding: A Comment on Dowton et al. Systematic Biology 63 (6): 1005-1009. DOI: 10.1093/sysbio/syu060
- Egloff W, Patterson D, Agosti D, Hagedorn G (2014) Open exchange of scientific knowledge and European copyright: The case of biodiversity information. ZooKeys 414: 109-135. DOI: 10.3897/zookeys.414.7717
- Guralnick R, Conlin T, Deck J, Stucky B, Cellinese N (2014) The Trouble with Triplets in Biodiversity Informatics: A Data-Driven Case against Current Identifier Practices. PLoS ONE 9 (12): e114069. DOI: 10.1371/journal.pone.0114069
- Guralnick R, Cellinese N, Deck J, Pyle R, Kunze J, Penev L, Walls R, Hagedorn G, Agosti D, Wieczorek J, Catapano T, Page R (2015) Community Next Steps for Making Globally Unique Identifiers Work for Biocollections Data. ZooKeys 494: 133-154. DOI: 10.3897/zookeys.494.9352
- Hinchliff C, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Cognill LM, Crandall KA, Deng J, Drew BT, Gazis R, Gude K, Hibbett DS, Katz LA, Laughinghouse IV H, McTavish EJ, Midford P, Owen CL, Ree R, Rees JA, Soltis DE, Williams T, Cranston KA (2015) Synthesis of phylogeny and taxonomy into a comprehensive tree of life. PNAS 112 (41): 12764-12769. DOI: 10.1073/pnas.1423041112
- Hugenholtz P, Tyson G (2008) Microbiology: Metagenomics. Nature 455 (7212): 481-483. DOI: 10.1038/455481a
- Kõljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AS, Bahram M, Bates S, Bruns T, Bengtsson-Palme J, Callaghan T, Douglas B, Drenkhan T, Eberhardt U, Dueñas M, Grebenc T, Griffith G, Hartmann M, Kirk P, Kohout P, Larsson E, Lindahl B, Lücking R, Martín M, Matheny PB, Nguyen N, Niskanen T, Oja J, Peay K, Peintner U, Peterson M, Põldmaa K, Saag L, Saar I, Schüßler A, Scott J, Senés C, Smith M, Suija A, Taylor DL, Telleria MT, Weiss M, Larsson K (2013) Towards a unified paradigm for sequence-based identification of fungi. Molecular Ecology 22 (21): 5271-5277. DOI: 10.1111/mec.12481
- Kraker P, Leony D, Reinhardt W, Gü NA, Beham n (2011) The case for an open science in technology enhanced learning. International Journal of Technology Enhanced Learning 3 (6): 643. DOI: 10.1504/ijtel.2011.045454
- Mietchen D (2014) The Transformative Nature of Transparency in Research Funding. PLoS Biology 12 (12): e1002027. DOI: 10.1371/journal.pbio.1002027
- Mietchen D, Hagedorn G, Willighagen E, Rico M, Gómez-Pérez A, Aibar E, Rafes K, Germain C, Dunning A, Pintscher L, Kinzler D (2015) Enabling Open Science: Wikidata

for Research (Wiki4R). Research Ideas and Outcomes 1: e7573. DOI: 10.3897/rio.1.e7573

- Miller J, Agosti D, Penev L, Sautter G, Georgiev T, Catapano T, Patterson D, King D, Pereira S, Vos R, Sierra S (2015) Integrating and visualizing primary data from prospective and legacy taxonomic literature. Biodiversity Data Journal 3: e5063. DOI: 10.3897/bdj.3.e5063
- Miller J, Dikow T, Agosti D, Sautter G, Catapano T, Penev L, Zhang Z, Pentcheff D, Pyle R, Blum S, Parr C, Freeland C, Garnett T, Ford LS, Muller B, Smith L, Strader G, Georgiev T, Bénichou L (2012) From taxonomic literature to cybertaxonomic content. BMC Biology 10 (1): 87. DOI: 10.1186/1741-7007-10-87
- Mindell D, Fisher B, Roopnarine P, Eisen J, Mace G, Page RM, Pyle R (2011) Aggregating, Tagging and Integrating Biodiversity Research. PLoS ONE 6 (8): e19491. DOI: 10.1371/journal.pone.0019491
- Mitraka E, Waagmeester A, Burgstaller-Muehlbacher S, Schriml LM, Su AI, Good BM (2015) Wikidata: A platform for data integration and dissemination for the life sciences and beyond. bioRxiv 0: 0. DOI: 10.1101/031971
- Nature (2015) Challenges in Irreproducible Research. http://www.nature.com/news/reproducibility-1.17552. Accession date: 2015 11 12.
- Neo4j (2015) From Relational to Neo4j. http://neo4j.com/developer/graph-db-vs-rdbms/. Accession date: 2015 11 16.
- Page RD (2014) The vision thing - it's all about the links. http://iphylo.blogspot.bg/2014/06/the-vision-thing-it-all-about-links.html. Accession date: 2015 12 02.
- Page RD (2015) Putting some bite into the Bouchout Declaration. http://iphylo.blogspot.bg/2015/05/putting-some-bite-into-bouchout.html. Accession date: 2015 12 02.
- Page RDM (2008) Biodiversity informatics: the challenge of linking data and the role of shared identifiers. Briefings in Bioinformatics 9 (5): 345-354. DOI: 10.1093/bib/bbn022
- Page RM (2006) Taxonomic names, metadata, and the Semantic Web. Biodiversity Informatics 3: 1-15. DOI: 10.17161/bi.v3i0.25
- Page RM (2011) Dark taxa: GenBank in a post-taxonomic world. http://iphylo.blogspot.bg/2011/04/dark-taxa-genbank-in-post-taxonomic.html. Accession date: 2015 11 12.
- Page RM (2015) Visualising big phylogenies (yet again). http://iphylo.blogspot.bg/2015/09/visualising-big-phylogenies-yet-again.html. Accession date: 2015 11 13.
- Pareja-Tobes P, Tobes R, Manrique M, Pareja E, Pareja-Tobes E (2015) Bio4j: a high-performance cloud-enabled graph-based data platform. bioRxiv 0: 0. DOI: 10.1101/016758
- Parr C, Guralnick R, Cellinese N, Page RM (2012) Evolutionary informatics: unifying knowledge about the diversity of life. Trends in Ecology & Evolution 27 (2): 94-103. DOI: 10.1016/j.tree.2011.11.001
- Patterson DJ, Cooper J, Kirk PM, Pyle RL, Remsen DP (2010) Names are key to the big new biology. Trends in Ecology & Evolution 25 (12): 686-691. DOI: 10.1016/j.tree.2010.09.004
- Penev L, Catapano T, Agosti D, Georgiev T, Sautter G, Stoev P (2012) Implementation of TaxPub, an NLM DTD extension for domain-specific markup in taxonomy, from the

experience of a biodiversity publisher. Journal Article Tag Suite Conference (JATS-Con) Proceedings 1: 1-14. URL: http://www.ncbi.nlm.nih.gov/books/NBK100351/

- Penev L, Lyal C, Weitzman A, Morse D, King D, Sautter G, Georgiev T, Morris R, Catapano T, Agosti D (2011) XML schemas and mark-up practices of taxonomic literature. ZooKeys 150: 89-116. DOI: 10.3897/zookeys.150.2213
- Penev L, Agosti D, Georgiev T, Catapano T, Miller J, Blagoderov V, Roberts D, Smith V, Brake I, Ryrcroft S, Scott B, Johnson N, Morris R, Sautter G, Chavan V, Robertson T, Remsen D, Stoev P, Parr C, Knapp S, Kress WJ, Thompson C, Erwin T (2010) Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. ZooKeys 50: 1-16.
- Pereira R, Richard K, Hobern D, Hyam R, Belbin L, Blum S (2009) TDWG Life Sciences Identifiers (LSID) Applicability Statement, Version 2009-09. Biodiversity Information. http://www.tdwg.org/standards/150. Accession date: 2015 11 03.
- pro-iBiosphere (2014a) Project Final Report. http://adm.pro-ibiosphere.eu/getatt.php?filename=oo_4751.pdf. Accession date: 2015 11 03.
- pro-iBiosphere (2014b) Open Biodiversity Knowledge Management System (OBKMS). http://adm.pro-ibiosphere.eu/getatt.php?filename=oo_4749.pdf. Accession date: 2015 11 03.
- Ratnasingham S, Hebert PN (2013) A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. PLoS ONE 8 (7): e66213. DOI: 10.1371/journal.pone.0066213
- Richards K (2010) TDWG GUID Applicability Statement, Version 2010-09. Biodiversity Information Standards (TDWG). http://www.tdwg.org/standards/150. Accession date: 2015 11 03.
- Shotton D (2009) Semantic publishing: the coming revolution in scientific journal publishing. Learned Publishing 22 (2): 85-94. DOI: 10.1087/2009202
- Smith MA, Fisher BL (2009) Invasions, DNA barcodes, and rapid biodiversity assessment using ants of Mauritius. Frontiers in Zoology 6 (1): 31. DOI: 10.1186/1742-9994-6-31
- Smith V, Georgiev T, Stoev P, Biserkov J, Miller J, Livermore L, Baker E, Mietchen D, Couvreur T, Mueller G, Dikow T, Helgen K, Frank J, Agosti D, Roberts D, Penev L (2013) Beyond dead trees: integrating the scientific process in the Biodiversity Data Journal. Biodiversity Data Journal 1: e995. DOI: 10.3897/bdj.1.e995
- van Godtsenhoven K, Elbaek MK, Sierman B, Bijsterbosch M, Hochstenbach P, Russell R, Vanderfeesten M (2009) Emerging Standards for Enhanced Publications and Repository Technology : Survey on Technology. Amsterdam University Press, 209 pp. URL: http://dx.doi.org/10.5117/9789089641892 DOI: 10.5117/9789089641892