Educational Technologies
Образователни технологии

IMPACT ON RESEARCH VISIBILITY USING STRUCTURED DATA AND SOCIAL MEDIA INTEGRATION

Nikolay Kasakliev, Elena Somova, Margarita Gocheva
University of Plovdiv "Paisii Hilendarski" (Bulgaria)

Abstract. The paper presents a study on applicability of structured data (metadata) in increasing of scientific articles' visibility, published on the Web. The usage of structured data in the description of the scientific content is discussed. The brief analyses of the used approaches for scientific publishing and some academic publishing systems are presented. The special attention is given to the metadata schemas (Metadata Object Description Schema, Dublin Core Metadata Element Set and Schema.org) used successfully for the publishing purposes. The appropriate integration of the information on scientific articles within social media (with Open Graph and Twitter Card tags) is also examined.

Keywords: structured data; scientific publishing; metadata schemas; academic publishing systems; social media integration

1. Introduction

In the era of globalization and the opening of scientific research, scientists are placed in a highly competitive environment. On one hand, the research units they work in must compete for project funding (Lepori, 2009), and the requirements for publishing the research in prestigious editions are rising on the other. In countries like Bulgaria, where the publication of research results is partly funded by the institutions in which scientists work, are encountered serious difficulties (Toshev, 2011). Very often, by financial reasons, results are published in lower ranked journals, journals that cease to exist in a few years, and those that are not indexed in prestigious bibliographic databases and archives. For these reasons, scientists also rely on publishing on personal or institutional websites where visibility of articles is weak.

In other situations, where publishing is made through academic publishing systems, there are limitations such as: lack of full text and meta-data level search, difficulty migrating to a newer version or to another publishing system, systems are not rendering mobile friendly content compatible with mobile devices' browsers (Google search using the mobile-friendliness of a website as part of the evaluation

in the search results (Schubert, 2016)), etc. (Eikebrokk, 2014) shows another problem with Open Access publishing systems, and it is using of PDF as a preferred publishing format and neglect formats like EPUB and MOBI. Another problem is confusion with names of the authors or article titles, because not all systems save all possible metadata (for example, information about organization affiliation of the author) or ORCID identifier. So, in these systems users cannot search by scientific area, subject domain or institutional affiliation.

All this shows that in order to increase the visibility of scientific papers, different approaches for publishing on the Web should be applied.

The goals of the paper are to make analysis of the used approaches of scientific publishing on the Web and examine impact on research visibility using metadata and social media integration. Section 2 presents applicability of Semantic Web and structured data in description of scientific articles. Common scientific publishing (on the Web) approaches are shown in Section 3. Section 4 introduces a brief study of academic publishing systems whereas Section 5 analyzes the most common metadata schemas. Section 6 points on social media integration of the scientific publications. Paper ends with Discussion section, presenting the importance of the study, and with Conclusion, giving summary of the authors' work.

2. Semantic Web and Structured Data

Until recently, most of the information on the Web was published in unstructured format (text, video, images or audio), which from the point of view of data has a number of limitations that concern mainly:

Searching information – It relied on keyword search, which allows search engines to be easily "misled" about the content of web pages;

- Retrieving information – Significant efforts were required from people to browse hundreds or thousands of web pages and retrieve the necessary information;
- Sharing information – It is not enough just to share a scientific result (a paper, an image, URL or other type). For example, for a scientific journal paper, the following data is very important: paper title, author(s), author organization affiliation, URL, paper abstract, key words, pages, journal name, journal volume, journal issue, publisher and publication date.

These limitations can be overcome by appropriate description (by metadata) of the content of scientific web pages. This approach, through adding metadata for the content, is called Semantic Web.

W3C[1] describes Semantic Web as "a common framework that allows data to be shared and reused across application, enterprise, and community boundaries". (Berners-Lee, 2001) defines Semantic Web as "a web of data that can be processed directly and indirectly by machines".

In the early 1960s the concept of the Semantic Network Model was presented as a form to represent semantically structured knowledge. These days this concept

extends the network of hyperlinked human-readable web pages by inserting machine-readable metadata about content of the pages and how they are related to each other. This enables automated agents (e.g. Web crawlers) to access the Web more intelligently.

The Semantic Web relies on formal ontologies (or vocabularies) to structure data for machine understanding. Ontology "formally describes concepts and relationships which can exist between them in some community. In other words, an ontology describes a part of the world." (Synak, 2009). An ontology provides a common way of representing knowledge in some domain.

At present, there are several languages for describing ontologies for Semantic Web: RDF, RDFS and OWL. Using these languages, we can add structure and meaning to the content of scientific web pages and link related data to one another.

Structured data is a standardized format for providing information about a web page and classifying the page content. Search engines like Google use structured data to "extract knowledge" from the content of the page. Social media such as Facebook use basic metadata structured properties for optimal display of the content. Structured data can be added to a web page using following formats: RDF, Microdata and JSON-LD.

3. Scientific Publishing on the Web

Scientific publishing on the Web is made mainly in: organization, team or personal websites; academic publishing systems.

The first approach is used in many situations. Here is some of them. The results of some research are not verified yet, but the author wants to make them visible to other scientists. Scientific paper is on draft stage and the author wants some feedback and recommendations. The author wants more visibility of the articles published somewhere else. Publishing articles that have been published on paper or in no longer existing online journal, proceeding, etc.

Publishing in organization, team or personal website has some advantages such as: easy and fast publishing, many document formats can be supported, opportunities for feedback, availability of adding more data about the paper, author, etc., personalization (customizing website structure and design, adding researcher custom content, integration of third party content, tracking page views), multi-language support (e.g. interface or search capabilities) and easy social media content integration. But this approach has also some limitations/lack of: advanced search (metadata-level search), larger costs of maintenance and upgrades, automation of export of a citation in different styles, website insights (statistics and reports), weak interoperability (needs manual export or import), features for collaboration with other researchers and indexing (academic databases and search engines).

The second approach is the basic way to publish scientific papers on the Web today. There are many reasons to publish on academic publishing systems, as (Marusic,

2014) points out "to increase the visibility of the journal in international indexing and citation databases, and ensure greater visibility in the global scientific community".

Publishing in academic publishing systems gives following advantages: easy publishing on low costs (pay only for publication fee), easy export of citation in different styles, generation of detailed statistics and reports for views and downloads, better indexing (in academic databases and search engines). But this approach has also some limitations/lack of: advanced search (metadata-level search), migration to both a newer software version or another software solution, interoperability (difficult import and export of content to different systems), multi-language support, support of different document formats and social media integration.

4. A Brief Study of Academic Publishing Systems

There are many software systems for scientific publishing. Almost all of them have similar features for submissions, peer review options, report builders and searching (by title, keyword, abstract, author, etc.). Publishers like Elsevier or Springer have their own publishing platforms, but they are not open and used by other publishers, that is why they are not analyzed in the paper.

We consider several academic publishing systems (see Table 1.), which are the most widespread, to examine their common features and whether the concept of the Semantic Web is implemented in these systems so they can ensure greater visibility of the research articles. The study is made only with a small amount of public information, since detailed information about the systems was not available to the public.

Table 1. shows some differences between examined publishing systems. Most of the systems are paid, so the publishers have to use sponsorship or publication fees to keep journals alive. Almost all of the systems use PDF's but only some of them support EPUB format, although EPUB is more suitable for online academic publishing because of its universal accessibility (Schwarz, 2018), responsiveness and its support of inline metadata, video and audio assets. Another major difference between systems is the ability for searching not only by simple keyword but by full text as well. No information on social media integration support was found for almost all systems.

Table 1. A brief study of academic publishing systems

Publishing system	Positives	Negatives
Arpha[2]	metadata import, authoring tool, semantic markup, journal statistics, online collaboration	no EPUB, paid, not open source, no altmetric data
eLife Libero[3]	workflow-based system, reach search capabilities, recommendations, metrics, open-source	PDF only, combination of many services and platforms/API, uses only JATS standard

Editoria[4]	authoring tool with track changes, version control, import functionality for DOCX files, open-source	no EPUB or PDF, no statistics, asset manager not available, book publication only (no journals), not full accessibility support
Fulcrum[5]	authors receive metrics of impact incl. altmetric scores, interoperable, EPUB 3.1 support, metadata-level search, open-source	currently is under development multimedia content oriented
Open journal system[6]	simple publishing procedure, over 30 languages supported, DOI support	no EPUB, Dublin Core meta-data only, no metadata-level search, indexing and biographical information is optional, no ORCID
Orvium[7]	open source, decentralized framework, full lifecycle traceability, Big Data analytics, online collaboration	uses own digital cryptocurrency for the payment of copyright licenses, no information about journal management tools, no information about supported document formats or authoring tools
Manifold[8]	multimedia files support, capabilities for converting old files, allows comments and annotations, full text search	uses only EPUB, Google Docs and HTML, no PDF and MS Word, no unique accounts, no metadata standard, Tweeter support only
Veruscript[9]	content indexing, data protection, analytics and reporting, DOI support, automated metadata capturing, plagiarism check	uses only PDF and HTML, hosted in cloud, paid

5. Metadata Schemas

Following the analysis of web publishing systems, we can conclude that in order to achieve a greater effect (visibility and the reputation of the author or academic journal) of publishing it is necessary to add additional data (metadata) to each scientific paper that can be easily interpreted by computers. National Information Standards Organization of the USA defines metadata as "structured information associated with an object for purposes of discovery, description, use, management and preservation" [10].

In the field of scientific publishing, we can categorize used metadata for identification and retrieval in three types:

- metadata describing a scientific article (like title, author, keywords, abstract, etc.);
- metadata describing a book, journal, etc., where the scientific articles are part of them (like title of journal/book, publishing date, publisher, volume, issue, etc.);
- metadata describing electronic resource (like name, versioning information, file format, and any other file technical information).

Metadata elements grouped into sets designed for a specific purpose, e.g., for a scientific publishing, are called metadata schemas. Metadata schema specifies the name and the semantics (meaning) of all used metadata elements in the schema. Many different metadata schemas are developed as standards across different domains, such as education, library, e-commerce, arts, etc. The paper presents three of them that are used successfully for the publishing purposes.

Metadata Object Description Schema (MODS)[11] is a schema with bibliographic element set, that can be used also in academic publishing systems. The MODS elements are divided into two levels: top elements (20 in number) and subelements, with attributes for elements and subelements. MODS defines top level elements like *titleInfo*, *note*, *name*, *subject*, *typeOfResource*, *classification*, *genre*, *abstract*, *part*, *tableOfContents*, etc. with many attributes which allow each scientific paper to been described in details. The schema is implemented in more than 35 digital libraries.

Dublin Core Metadata Element Set (DC) contains a small set of 15 basic properties for general purposes usage[12]. DC has some weaknesses and loses its popularity in digital publishing. The major weakness is its simplicity which results in a loss of specificity that lead to low interoperability (difficult conversion of DC into/from other systems) (Beall, 2004). (Gartner, 2003) defines that the dual approach to semantic breadth (the use of simple fields and qualifiers to refine them) that DC has taken has reduced its value as either a metadata container or as a medium of exchange. The 15 elements are often found to be too broad. For example, the DC field *creator* can cover a large array of individuals or organizations responsible for the creation of an object (like authors, editors, compilers, etc.). Concatenating all of these data into one field is often inadequate for a useful bibliographic description. On the other hand, qualifying elements to distinguish them (for example, authors from editors) in the *creator* field reduces the interoperability of the DC record.

Schema.org was launched by the major search engines Bing, Google, and Yahoo in 2011, providing a single schema across a wide range of topics that included people, places, events, products, offers, etc. (Guha, 2016). Nowadays it is used in many applications like Google's Gmail and Search, Microsoft's Cortana, Yandex, Apple's Siri, etc. The main advantages of this schema are extensibility, mass adoption and covering the most common use cases (Khalili, 2013). It supports many types, but *ScholarlyArticle* type can be used for scientific publishing. This type specifies that associated content is written by experts in academic or professional fields and gives information about what has been studied or researched on a topic.

For digital publishing schema.org can be successfully integrated with RDF, Microdata and JSON-LD. Microdata is integrated easy with HTML. Microdata was introduced in HTML5 and allows the author of an EPUB document to insert supporting vocabularies along with name-value pair in an existing markup nested content. The vocabularies give information to the search engine about the content,

the content type (book, movie, person, event, article, etc.), etc. and help the search engine to understand better the underlying meaning of the content.

6. Social Media dnd Scientific Publishing

Social media (incl. social networks) is increasingly influencing by academic life, not only in communication, but also in spreading research results and scientific publications. Factors such as comments, citations and the number of social media posts have already been taken into account when evaluating researchers' scientific achievements (Kasakliev, 2016). That is why researchers look at possibilities to share their work on social media as way to increase the visibility of the publications. For that reason, it is not enough just to share the simple link (URL) to the scientific content, but to add information like title, URL, image, audio/video, short description, etc.

Open Graph protocol was originally created at Facebook and is inspired by Dublin Core, link-rel canonical, Microformats and RDFa[13]. While many different technologies and schemas exist and could be combined together, there is not a single technology which provides enough information to richly represent any web page within the social graph (Haugen, 2010). This protocol relies on adding basic metadata (with *<meta>* tags in the *<head>*) to a web page. There are four required properties (*og:title*, *og:type*, *og:image* and *og:url*), but some optional properties also can be used (such as *og:audio*, *og:description*, *og:locale* and *og:video*). We suggest for describing research articles, developers to use *article*, *book* and *website* values for *og:type*. Some properties can have extra metadata (structured) attached to them. Open Graph can be used along with other markup such as Twitter Cards.

Twitter Card tags are similar to Open Graph tags, and are based on the Open Graph protocol[14]. It is easy to been used by adding proper HTML markup to the *<head>* section of the web page: *<meta name="twitter:card" content="summary"> </meta>*. Twitter Card allows easy overcoming 140 characters' limitation for the posts and gives authors more visibility of their research.

7. Discussion

For each scientist, it is of particular importance that the results of his/her research be verified, multiplied or further developed. To achieve this, they need to be easily accessible to the widest possible range of people. This can be done by adding semantics (metadata) to the published digital content, which leads to several benefits:

– possibility of semantic search, retrieve and spread of scientific information;
– more machine-readable scientific articles with intelligent assistants (Siri, Cortana, Alexa, etc.), search engines and e-books readers;
– increase the visibility to community (reputation);
– easy identification of related research papers or journals;

– long-term preservation and management of scientific information;
– no need to build search features into the personal or journal web pages, instead of using services like Google Custom search;
– richer content with files in formats like EPUB, where different multimedia elements can be included;
– sharing, in optimal way, the research articles or journals information in social media.

This approach needs to continue to be carefully explored, but it can be assumed that the visibility of scientific research can be significantly increased.

8. Conclusion

The paper presented a brief study of used approaches in scientific publishing and gave some recommendations to increase the visibility of the authors' research such as: adding more metadata for description of the scientific content, applying search engines' optimization techniques, migrating existing web pages with scientific papers to the Semantic Web 3.0, promoting academic content in Internet and sharing scientific results on social and scientific networks.

The use of Semantic Web with its formal ontologies (vocabularies) and structured data was explained for description of the scientific content. The most common metadata schemas were discussed as well as the appropriate way of integration of the scientific papers within the social media (networks).

We studied the publishing approach on an organizational or personal web page and pointed out some main disadvantages like: missing of metadata level search, inappropriate sharing on social media, interoperability and generation of citation texts in different styles.

A brief analysis of academic publishing systems was proposed in order to help in the choice of such system. Some limitations of publishing systems have been identified, such as: not optimal social media integration, lack of full text search, missing of unique identifiers (e.g. DOI or ORCID), interoperability, migration to both a newer version and another software solution and difficult import and export to several e-documents formats.

NOTES

1. W3C Semantic Web, http://www.w3.org/2001/sw/
2 Arpha, https://arphahub.com/about/platform
3. Libero, https://libero.pub/products/
4. Editoria, https://editoria.pub/
5. Fulcrum, https://www.fulcrum.org/
6. Open Journal Systems, https://pkp.sfu.ca/ojs/

7. Orvium, https://orvium.io/
8. Manifold, https://manifoldapp.org/
9. Veruscript, https://www.veruscript.com/
10. A Framework of Guidance for Building Good Digital Collections, NISO, https://www.niso.org/sites/default/files/2017-08/framework3.pdf
11. Metadata Object Description Schema, http://www.loc.gov/standards/mods/
12. Dublin Core, http://dublincore.org/documents/dces/
13. Open Graph protocol, https://ogp.me/
14. Optimize Tweets with Cards, https://developer.twitter.com/en/docs/tweets/optimize-with-cards/guides/getting-started.html

REFERENCES

Lepori, B. et al., (2009). Comparing the organization of public research funding in central and eastern European countries. *Science and Public Policy*, v. 36, 9, pp. 667 – 681, DOI: 10.3152/030234209X479494.

Toshev, B. (2011). Status and problems of the Bulgarian Science Fund (1990 – 2011). *Bulgarian Journal of Science and Education Policy (BJSEP),* v. 5, 1. [in Bulgarian]

Schubert, D. (2016). Influence of mobile-friendly design to search results on Google search. *19th International Conference Enterprise and Competitive Environment, 10 – 11 March 2016, Brno, Czech Republic*, 424 – 433. doi.org/10.1016/j.sbspro.2016.05.517.

Eikebrokk, T. et al. (2014). EPUB as publication format in Open Access journals: Tools and workflow. *Code 4 Lib Journal*, Issue 24, ISSN 1940-5758.

Berners-Lee, T. et al. (2001). The Semantic Web in Scientific American. *Scientific American.* Vol. 284, No. 5, 34 – 43. https://www.researchgate.net/publication/243773883

Synak, M., Dabrowski, M. & Kruk, S. R. (2009). Semantic Web and Ontologies. In: *Kruk S.R., McDaniel B. (eds) Semantic Digital Libraries*. Berlin, Heidelberg: Springer, https://doi.org/10.1007/978-3-540-85434-0_3

Marusic, A. (2014). Publishing scientific journals in the digital age: opportunities for small scholarly journals. *Prilozi*, Volume 35, Issue 3, 17 – 21. https://doi.org/10.1515/prilozi-2015-0003

Schwarz, T. et al. (2018). Accessible EPUB: Making EPUB 3 Documents Universal Accessible. Computers helping people with special needs, *16th International Conference, ICCHP 2018*, Linz, Austria.

Beall, J. (2004). Dublin Core: An Obituary, *Library Hi Tech News*, (vol. 21), Issue: 8, 40 – 41, https://doi.org/10.1108/07419050410567399.

Gartner, R. (2003). *MODS: Metadata Object Description Schema*. Pearson New Media Librarian Oxford University Library Services.

Guha, R. et al. (2016) Schema. org: Evolution of structured data on the web. *Communications of the ACM*, (vol. 59) No. 2, 44 – 51 doi: 10.1145/2844544.

Khalili, A. & Auer, S. (2013). WYSIWYM Authoring of structured content based on Schema.org. *Web Information Systems Engineering – WISE 2013*, 425 – 438.

Kasakliev, N. & M. Atanasova (2016). Web based citation manager. *Journal of the Sofia University for Educational Research*, (vol. 1), 35 – 48, ISSN 1314-8753, [in Bulgarian].

Haugen, A. (2010). Abstract: The open graph protocol design decisions. *International Semantic Web Conference*, The Semantic Web – ISWC 2010, 338 – 338.

✉ **Dr. Nikolay Kasakliev, Assoc. Prof.**
ORCID iD 0000-0003-4010-144X,
Researcher ID: P-6676-2019,
Scopus Author ID: 34879939800
Department "Computer Science"
University of Plovdiv "Paisii Hilendarski"
4000 Plovdiv, Bulgaria
E-mail: kasakliev@uni-plovdiv.bg

✉ **Dr. Elena Somova, Assoc. Prof.**
ORCID iD: 0000-0003-3393-1058;
Researcher ID: P-6765-2019;
Scopus Author ID: 35932836600
Head of Department "Computer Science"
University of Plovdiv "Paisii Hilendarski"
4000 Plovdiv, Bulgaria
E-mail: eledel@uni-plovdiv.com

✉ **Ms. Margarita Gocheva, PhD student**
ORCID ID: 0000-0002-7739-5915,
University of Plovdiv "Paisii Hilendarski"
4000 Plovdiv, Bulgaria
E-mail: gocheva@au-plovdiv.bg