

# EPUB3 for Integrated and Customizable Representation of a Scientific Publication and its Associated Resources

Hajar Ghaem Sigarchian<sup>1</sup>, Ben De Meester<sup>1</sup>, Tom De Nies<sup>1</sup>, Ruben Verborgh<sup>1</sup>, Wesley De Neve<sup>1,2</sup>, Erik Mannens<sup>1</sup>, and Rik Van de Walle<sup>1</sup>

<sup>1</sup> Ghent University - iMinds - Multimedia Lab

Gaston Crommenlaan 8 bus 201, B-9050 Ledeborg-Ghent, Belgium

<sup>2</sup> Korea Advanced Institute of Science and Technology (KAIST) - IVY Lab  
Yuseong-gu, Daejeon, Republic of Korea

{hajar.ghaemsigarchian, ben.demeester, tom.denies, ruben.verborgh, wesley.deneve, erik.mannens, rik.vandewalle}@ugent.be

**Abstract.** Scientific publications point to many associated resources, including videos, prototypes, slides, and datasets. However, discovering and accessing these resources is not always straightforward: links could be broken, readers may be offline, or the number of associated resources might make it difficult to keep track of the viewing order. In this paper, we explore potential integration of such resources into the digital version of a scientific publication. Specifically, we evaluate the most common scientific publication formats in terms of their capability to implement the desirable attributes of an enhanced publication and to meet the functional goals of an enhanced publication information system: PDF, HTML, EPUB2, and EPUB3. In addition, we present an EPUB3 version of an exemplary publication in the field of computer science, integrating and interlinking an explanatory video and an interactive prototype. Finally, we introduce a demonstrator that is capable of outputting customized scientific publications in EPUB3. By making use of EPUB3 to create an integrated and customizable representation of a scientific publication and its associated resources, we believe that we are able to augment the reading experience of scholarly publications, and thus the effectiveness of scientific communication.

## 1 Introduction

Scientific publications consist of more than only text: they may also point to many associated (binary) resources, including videos, prototypes, slides, and datasets. Yet today, only the access to the *text* of a scientific publication is straightforward; the associated resources are often more difficult to access. For instance, readers may not always have an Internet connection at their disposal to download related materials, and even when this is the case, links might become broken after a while. Furthermore, given their diverse nature, related materials often need to be accessed in a different reading environment like a standalone media player, causing readers to lose track of the scientific narrative.

The 2007 Brussels Declaration<sup>3</sup> by the International Association of Scientific, Technical and Medical (STM) Publishers states that “raw research data should be made freely available” and that “one size fits all solutions will not work”. In this paper, we illustrate that the ability to (adaptively) create an integrated representation of a scientific publication and its associated resources contributes to these goals. Specifically, we evaluate the most common scientific publication formats in terms of their capability to implement the desirable attributes of an enhanced publication and to meet the functional goals of an enhanced publication information system: PDF, HTML, EPUB2, and EPUB3. In addition, we present an EPUB3 version of an exemplary publication in the field of computer science, integrating and interlinking an explanatory video and an interactive prototype. Finally, we introduce a demonstrator that is capable of outputting customized scientific publications in EPUB3.

The rest of this paper is structured as follows. In Section 2, we discuss a number of current best practices among three scientific publishers, focusing on the way open formats and their features are used to enhance scientific publications. Next, in Section 3, we investigate to what extent PDF, HTML, EPUB2, and EPUB3 facilitate the use of enhanced scientific publications and corresponding information systems. In Section 4, we present an exemplary scientific publication in EPUB3 that integrates an explanatory video and an interactive prototype. In Section 5, we introduce our demonstrator for creating customized scientific publications in EPUB3. Finally, in Section 6, we present our conclusions and a number of directions for future work.

## 2 Current Best Practices

In this section, we briefly discuss a number of current best practices among three scientific publishers, focusing on the way open formats are used to make available scientific publications that have been enhanced with multimedia, interactivity, and/or Semantic Web features.

**BioMed Central and Hindawi Publishing Corporation:** These publishers make scientific publications available in several formats: PDF, HTML, and EPUB2. The HTML version of the publications can for instance be enhanced with reusable data (e.g., supplementary datasets), while the EPUB2 version of the publications just uses links to cited publications in EPUB2 format. However, the publications in question do not contain any embedded interactive multimedia content.

**Elsevier:** Elsevier makes available different versions of a scientific publication: PDF, HTML, MOBI, and EPUB2. In addition, authors are able to deposit their datasets, making it possible for readers to access and download these datasets [1]. Moreover, the EPUB2 version of a publication is enriched with direct links to the PDF version of cited publications, thus not embedding these PDF versions into the EPUB2 file. Furthermore, the EPUB2 version of a publication does not contain any embedded interactive multimedia content.

---

<sup>3</sup> <http://www.stm-assoc.org/brussels-declaration/>

In summary, we can conclude that none of the aforementioned EPUB2 versions – as currently made available by BioMed Central, Hindawi Publishing Corporation, and Elsevier – embed interactive multimedia content for offline usage (i.e., readers need to have network connectivity in order to be able to access all linked resources), nor do they contain Semantic Web features.

### 3 Comparative Analysis of Publication Formats

In recent years, a new open format for distribution and interchange of digital publications has emerged, called EPUB3 [6]. This format can also be used in the context of scientific publications. In what follows, we investigate to what extent PDF, HTML, EPUB2, and EPUB3 are able to support the properties of an enhanced scientific publication (that is, a scientific publication with multimedia, interactivity, and/or Semantic Web features). To that end, we analyzed a number of desirable attributes of an enhanced publication. Furthermore, we also investigated the functional goals of an enhanced publication information system (that is, the system that facilitates the authoring of enhanced publications).

Thoma *et al.* [10] defined a core set of nine desirable attributes of an enhanced publication: *appearance, page transitions, in-page navigation, image browsing, navigation to an embedded/linked media object, support for interactivity, transmission, embedding and linking of multimedia/interactive objects, and document integrity and structure*. In addition, by both considering the attributes defined by Thoma *et al.* in [10] and a review of five already existing enhanced publications, Adriaansen *et al.* [2] identified eleven attributes of an enhanced publication: *navigation by table of contents, metadata, links to figures and tables, attached data resources, link from text to references, direct publication links from references, reader comments, download as PDF, interactive content, relations, and cited by*. Furthermore, as argued in a talk by Ivan Herman<sup>4</sup>, *bridging online and offline access* is a need for high-quality digital books, and consequently for high-quality digital scientific publications, given that offline access enables users to access supplementary information, even when they do not have a network connection at their disposal. As a result, although none of the aforementioned research efforts discusses this aspect, we consider offline access to be a desirable attribute of an enhanced publication as well.

Besides the attributes of enhanced publications, we also considered data model and information system aspects. Bardi *et al.* [3] reviewed existing data models for enhanced publications, taking into account structural and semantic features, also proposing a classification scheme for enhanced publication information systems based on their main functional goals. In this context, the authors outline four major scientific motivations that explain the functional goals of an enhanced publication information system: *packaging with supplementary material, improving readability and understanding, interlinking with research data, and enabling repetition of experiments*. Furthermore, we believe that *portability*

---

<sup>4</sup> <http://www.w3.org/2014/Talks/0411-Seoul-IH/Talk.pdf>

is also needed in order to preserve the availability of resources and their inter-linking, given that it enables users to even access supplementary information in offline situations. Thus, an enhanced publication that has supplementary resources needs to be a self-contained package. Therefore, we identified *portable packaged file* as another desirable attribute of an enhanced publication.

Finally, according to Liu [8], users are in need of a hybrid solution for print and digital resources. This means that, besides all different digital publication formats, *print* also remains an important publication medium. As a result, we see *suitable for print* as another desirable attribute of an enhanced publication.

Ideally, an enhanced publication information system should be able to support all the desirable attributes mentioned above. Considering the desirable attributes of enhanced publications and the functional goals of enhanced publication information systems, we mapped the attributes identified in [10,2] onto each functional goal identified by Bardi *et al.* in [3]. Our mapping can be found in the first and second column of Table 1. We can observe that nearly all desirable attributes of an enhanced publication can be covered by the functional goals of an enhanced publication information system, with the exception of the final three attributes, for which we defined our own functional goals.

Next, we investigated what scientific publication formats are the most promising to cover both the desirable attributes of an enhanced publication and the functional goals of an enhanced publication information system. We have summarized our findings in the four rightmost columns of Table 1. Corresponding explanatory notes can be found below.

**Packaging with supplementary material:** This functional goal states that it should be possible to add supplementary material to a scientific publication. PDF can embed audio and video but it does not support rich media (e.g., media overlays). As such, it is not a suitable format for embedding various types of associated resources (e.g., interactive content and standalone applications). Consequently, PDF has limited support for this functional goal and its underlying attributes. Note that extensions exist, such as export to a PDF Portfolio in Adobe Acrobat<sup>5</sup>, that make it possible to combine related materials. However, to the best of our knowledge, none of these extensions for instance allow embedding interactive content and standalone applications. Furthermore, the embedded resources are not reusable, unlike the EPUB3 format, which lets users reuse embedded resources. In order to package research data within an HTML file, all the dependencies need to be packaged as well. While this is possible (e.g., using a zipped folder), there is no standardized approach to do this, as opposed to EPUB2 and EPUB3. Therefore, we do not consider HTML to be suitable for meeting this functional goal. According to the EPUB2 specification [7], EPUB2 cannot embed multimedia and interactive objects. Consequently, EPUB2 also offers limited support for this functional goal. However, in EPUB3, no such restrictions are specified. As a result, we can conclude that EPUB3 is the only format that fully supports this functional goal.

---

<sup>5</sup> <http://www.adobe.com/products/acrobat/combine-pdf-files-portfolio.html>

Functional Goals	Attributes	Format			
		PDF	HTML	EPUB2	EPUB3
Packaging with supplementary material	<ul style="list-style-type: none"> <li>– Embedding and linking of multimedia/interactive objects</li> <li>– Document integrity and structure</li> <li>– Attached data resources</li> <li>– Navigating to an embedded / linked media object</li> </ul>	✓*			✓
Enabling repetition of experiments	<ul style="list-style-type: none"> <li>– Native support for interactivity</li> <li>– Code execution</li> <li>– Interactive content</li> </ul>		✓		✓
Improving readability and understanding	<ul style="list-style-type: none"> <li>– Navigation by table of contents</li> <li>– Reader comments</li> <li>– Appearance</li> <li>– Page transitions</li> <li>– In-page navigation</li> <li>– Image browsing</li> <li>– Links to figures and tables</li> <li>– Direct publication links from references</li> <li>– Cited by</li> </ul>	✓*	✓*	✓	✓
Interlinking with research data	<ul style="list-style-type: none"> <li>– Metadata</li> <li>– Relations</li> </ul>		✓	✓*	✓
Portable packaged file	<ul style="list-style-type: none"> <li>– Bridging online / offline</li> <li>– Transmission</li> </ul>	✓*		✓*	✓
Suitable for print	<ul style="list-style-type: none"> <li>– Download as PDF</li> </ul>	✓			

Table 1: Support for enhanced publication attributes (\* = limited support).

**Enabling repetition of experiments:** This functional goal aims at enabling researchers to (re-)execute experiments and/or demonstrators from within a scientific publication. PDF has limited support for scripting and code execution. However, the support available is not sufficient for building small standalone applications that can act as interactive content (e.g., self-contained widgets). As a result, PDF is not suitable for meeting this functional goal. HTML is able to embed code (e.g., JavaScript). Moreover, thanks to the inline frame element (that is, the `iframe` element), HTML can also be used as an interface to other experiments. As EPUB2 does not support JavaScript, it is not suited for repetition of experiments. However, similar to HTML, EPUB3 supports JavaScript, and thus the aforementioned functional goal

(unless experiments are involved that for instance use complex algorithms on clusters to obtain their results).

**Improving readability and understanding:** PDF is a specific format for print, and not for screen readers. While still undeniably the most suitable format for print layout, in digital form, it does not have device independence [5], making it difficult to maintain readability on different screens. According to the PDF specification, it has a limited support for this functional goal. On the other hand, HTML, EPUB2, and EPUB3 are suitable for improving readability and understanding, because they can overcome the aforementioned shortcomings of PDF (*cf.* the use of reflowable layout).

**Interlinking with research data:** In order to make links between supplementary materials added to publications, (relational) metadata need to be taken into account. PDF has a coarse level of support for metadata (e.g., title and author information), and where these metadata are not related to interlinking supplementary materials. As a result, PDF is not suitable for meeting this functional goal. HTML can be enriched for interlinking purposes using Semantic Web formats and technologies [9] (e.g., RDF and OWL). EPUB2 has limited support for metadata. Furthermore, it does not allow embedding multimedia and interactive content as supplementary research data. Hence, EPUB2 is not suitable for meeting this functional goal. According to the EPUB3 specification, it supports metadata and interlinking of research data. In fact, it retains all functionality of (X)HTML5.

Apart from a suitable format, interlinking supplementary materials requires suitable ontologies. Fortunately, many suitable candidates for general and specific interlinking purposes are already available. For example, `schema.org` is an ontology that is suitable for use in a variety of domains, including the description of events and creative works. It can thus be used to semantically enhance publications, and it can also be extended by other ontologies. Furthermore, Standard Analytics<sup>6</sup> aims at turning scholarly publications into an interface to a web of data, making use of already existing web ontologies. Moreover, Structural, Descriptive, and Referential (SDR)<sup>7</sup> is an ontology for representing academic publications, related artifacts (e.g., videos, slides, and datasets), and referential metadata. This ontology can generically define all possible interactive and multimedia resources. In addition, any publication can use general ontologies such as the Citation Typing Ontology (CiTO)<sup>8</sup>, the Bibliographic Ontology (BIBO)<sup>9</sup>, and the Common European Research Information Format (CERIF)<sup>10</sup>. Finally, publications may also need to make use of ontologies that are specific for their research domains (e.g., in the medical domain, the Infectious Disease Ontology (IDO)<sup>11</sup> could be used).

<sup>6</sup> <https://standardanalytics.io/>

<sup>7</sup> <http://onlinelibrary.wiley.com/doi/10.1002/asi.23007/full>

<sup>8</sup> <http://www.essepuntato.it/lode/http://purl.org/spar/cito>

<sup>9</sup> <http://bibliontology.com/>

<sup>10</sup> [http://helios-eie.ekt.gr/EIE/bitstream/10442/13864/1/IJMS0\\_2014\\_CERIF\\_authorFinalVersion.pdf](http://helios-eie.ekt.gr/EIE/bitstream/10442/13864/1/IJMS0_2014_CERIF_authorFinalVersion.pdf)

<sup>11</sup> [http://infectiousdiseaseontology.org/page/Main\\_Page](http://infectiousdiseaseontology.org/page/Main_Page)

**Portable packaged file:** PDF has limited support for packaging interactive content and standalone applications. Furthermore, it cannot bridge the gap between online and offline usage. Indeed, PDF is an offline format for print, and any interactive parts will not remain after printing a publication. As mentioned before, HTML lacks a proper packaging structure, making this format not a suitable candidate for meeting this functional goal. A similar remark holds regarding EPUB2, as this format does not have support for embedding interactive multimedia resources. As EPUB3 has extensive support for embedding interactive multimedia resources, it can be considered a suitable format for creating portable packaged files. Ideally, users expect that all types of resources can be embedded in a packaged file, regardless of their size. This is one of the shortcomings of EPUB3. Embedding large datasets makes the size of an EPUB3 file potentially very large, causing portability and readability issues. We discuss a possible solution to this issue in Section 5.

**Suitable for print:** Currently, PDF is the only format suitable for print. Although HTML, EPUB2, and EPUB3 can also be used for the purpose of print, they have been designed for screen readers and can currently not match the high typesetting demands for print publications.

As can be seen in Table 1, EPUB3 is the format that supports most desirable attributes of an enhanced publication and most functional goals of an enhanced publication information system. Only PDF is suitable for print output, given that HTML and EPUB(2/3) have been primarily designed for screen output, typically resulting in a layout that is suboptimal for print. Note that, as a workaround for this problem, the EPUB(2/3) and HTML versions of a publication can embed or link to the PDF version of a publication.

## 4 Proof-of-Concept: A Scientific Publication in EPUB3

In this section, we demonstrate how EPUB3 can be used to create an integrated representation of a scientific publication and its associated resources. To that end, we enhanced the “Everything is Connected” publication [11] – a paper authored by ourselves and a number of colleagues – embedding an explanatory video and an interactive prototype. The resulting proof-of-concept is available for download<sup>12</sup>. We used Radium<sup>13</sup> as our electronic reading system, since it supports most features of EPUB3. As illustrated by Figure 1, our proof-of-concept shows how a publication can act as an interface to different types of research outputs. Note that, instead of adding a link to the online version of the interactive prototype, we made use of an `iframe` to allow immediate access to the interactive prototype from within the publication, thus not requiring the reader to make use of a different reading environment.

<sup>12</sup> <http://multimedialab.elis.ugent.be/users/hghaemsi/EnhancedPublication.epub>

<sup>13</sup> <http://readium.org/>

Furthermore, we semantically enhanced our exemplary EPUB3 publication by making use of [schema.org](http://schema.org), a general ontology that allows describing books and articles, among other creative works. Thanks to properties such as `embedUrl`, `description`, and `contentUrl`, [schema.org](http://schema.org) makes it possible to indicate how a resource is related to the target EPUB3 publication in a straightforward way. We illustrate this in Figure 2. Note that [schema.org](http://schema.org) is supported by major search engines such as Bing, Google, Yahoo!, and Yandex. However, at the time of writing this paper, the aforementioned search engines did not have support yet for indexing EPUB3 publications (and reading the metadata available within these publications).

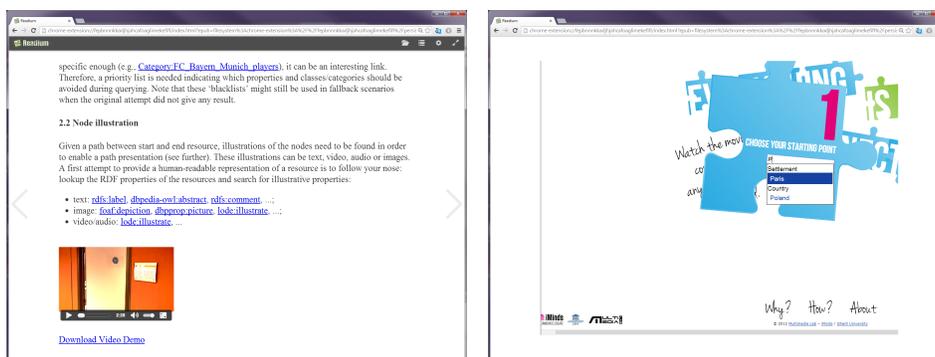


Fig. 1: Exemplary scientific publication enhanced with an explanatory video (left) and an interactive prototype (right). Both the video and the prototype have been embedded into the EPUB3 version of the scientific publication.

```
<div vocab="http://schema.org/" property="video" typeof="VideoObject">
  <p>Below, you can find <span property="Description">an embedded screencast</span>.</p>
  <video width="320" height="240" controls="">
    <source property="embedUrl" src="/video.mp4" type="video/mp4"/>
  </video>
  <p>This screencast can also be accessed <a property="url" href="http://youtu.be/FavygFT5Brs">remotely</a>,
  or can be <a property="contentUrl" href="/video.mp4">downloaded</a>.</p>
</div>
```

Fig. 2: Use of [schema.org](http://schema.org) for interlinking a local and remote video object.

## 5 Creating Customized EPUB3 Publications

In the previous sections, we explained how supplementary materials can be embedded into a scientific publication. As mentioned before, embedding all relevant

supplementary materials in a portable packaged file is not always cost-effective and/or desirable for a user. Since the size of an EPUB3 file is dependent on the size of all embedded resources, it will not be lightweight in all use cases, e.g., when embedding large datasets. The problem is that, on the one hand, a packaged file should not face portability and other usage issues relevant to its size. On the other hand, the advantages of having a portable packaged publication are overthrown with the disadvantage of not being able to distribute the entire publication properly. Users may not need all embedded supplementary materials and instead, wish to have their own customized lightweight publication. For instance, we can refer to big datasets or high-resolution images which can be located in a remote repository instead of embedding them in the portable packaged file. An environment for outputting customized publications allows users to select and embed the supplementary materials to the extent that they choose. Hence, they can determine the size of the EPUB3 file themselves. That way, the problem of distributing overly large publications is solved, and only the content that the user needs is distributed. The only disadvantage of this approach is the added complexity at the distribution side (i.e., at the platform of the publisher). However, most publishers already have an extensive online distribution infrastructure, which could easily be expanded with an interface such as the one we propose. For example, publishers such as Elsevier offer different formats of a publication to users. In particular, on the ScienceDirect website of Elsevier, there is an option for the user to select his/her preferred format.

To illustrate this concept of customizable publications, we implemented a basic demonstrator in which a user can first select the relevant supplementary material using a web interface, after which a customized EPUB3 publication is outputted. Figure 3 shows the user interface of our online demonstrator. Content selection is entirely done at the client side, based on the HTML representation of a publication. The selected content is then packaged as an EPUB3 file on the server side. The resulting demonstrator is available online<sup>14</sup>. Note that the author of a publication can determine which elements are customizable, simply by adding the class `customizable` to the desired HTML elements.

Ideally, the implemented functionality for outputting customized publications in EPUB3 would be integrated into an authoring environment, where authors and publishers could indicate which elements of a publication are customizable. In previous work, we have implemented such an authoring environment for the collaborative creation of enriched e-Books using EPUB3 [4]. It allows authors and publishers to create an electronic publication with all required material embedded. Next, this publication can be exported as an EPUB3 file. In future work, we aim to showcase an integrated version of this authoring environment with a customizable distribution platform as described above.

---

<sup>14</sup> <http://uvdt.test.iminds.be/custompublication/books/1/main.xhtml>

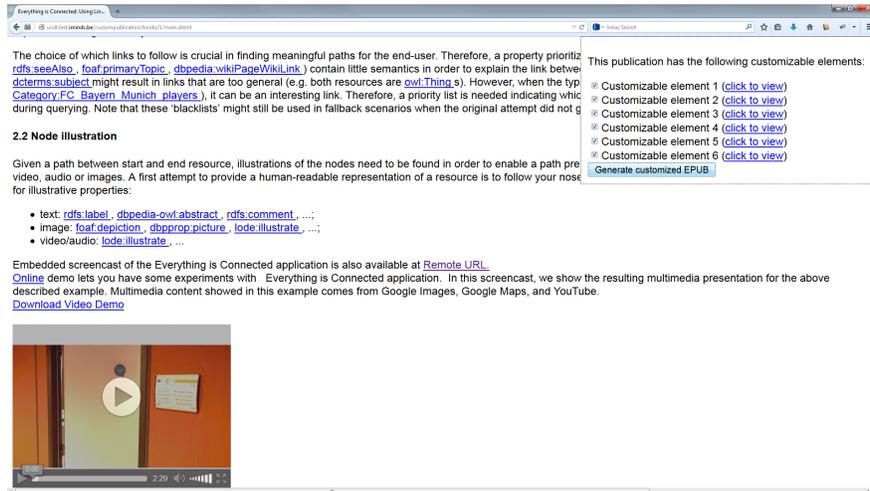


Fig. 3: The interface of our demonstrator for creating customized publications. Users can select the supplementary materials that they want to have embedded in the EPUB3 version of the enhanced publication.

## 6 Conclusions and Future Work

In this paper, we demonstrated that the increasingly popular EPUB3 format can be used to create integrated representations of a scientific publication and its associated resources. By doing so, we believe that this contributes to a better reading experience and more effective scientific communication (e.g., support for the inclusion of explanatory videos and interactive prototypes should enable authors to better transfer their knowledge and experience). In addition, we indicated that an EPUB3 version of a scientific publication can be used as a primary version, from which other versions of the scientific publication can be reached (e.g., a PDF version for print), thereby allowing legacy content to persist.

We can identify a number of directions for future research. First, user-friendly authoring tools are needed that allow easily creating enhanced scientific publications, and where these scientific publications can act as an interface to different research outputs. We have already started taking steps in this direction. Second, these authoring tools need to support different output formats, in order to meet the needs of both readers that are reading on paper and readers that are reading digitally. Third, these authoring tools also need to make it possible to easily add metadata to EPUB3 versions of scientific publications, such that EPUB3 versions of scientific papers may have the same degree of discoverability as PDF and HTML versions. Finally, it would be interesting to investigate the good practices of novel publication repositories such as PLOS ONE, Figshare, and ResearchGate.

## 7 Acknowledgments

The research activities described in this paper were funded by Ghent University, iMinds (a research institute founded by the Flemish Government), the Institute for Promotion of Innovation by Science and Technology in Flanders (IWT), the FWO-Flanders, and the European Union.

## References

1. Aalbersberg, I.J., Dunham, J., Koers, H.: Connecting Scientific Articles with Research Data: New Directions in Online Scholarly Publishing. *Data Science Journal* 12(0), WDS235–WDS242 (2013)
2. Adriaansen, D., Hooft, J.: Properties of Enhanced Publications and the Supporting Tools
3. Bardi, A., Manghi, P.: Enhanced Publications: Data Models and Information Systems. *LIBER Quarterly* 22 (2014)
4. De Meester, B., De Nies, T., Ghaem Sigarchian, H., Vander Sande, M., Van Campen, J., Van Impe, B., De Neve, W., Mannens, E., Van de Walle, R.: A Digital-First Authoring Environment for Enriched e-Books using EPUB 3. In: Proceedings of the 18th Int'l. Conference on Electronic Publishing (ELPUB), June 19-20, Thessaloniki, Greece (2014)
5. Eikebrokk, T., Dahl, T.A., Kessel, S.: EPUB as Publication Format in Open Access Journals: Tools and Workflow. *Code4Lib Journal* 24 (2014)
6. IDPF: Electronic Publication, version 3. <http://idpf.org/epub/30>
7. IDPF: Open Publication Structure (OPS). [http://www.idpf.org/epub/20/spec/OPS\\_2.0.1\\_draft.htm#Section1.3.7](http://www.idpf.org/epub/20/spec/OPS_2.0.1_draft.htm#Section1.3.7)
8. Liu, Z.: Print vs. Electronic Resources: A Study of User Perceptions, Preferences, and Use. *Information Processing & Management* 42(2), 583–592 (2006)
9. Shotton, D.: Semantic Publishing: The Coming Revolution in Scientific Journal Publishing. *Learned Publishing* 22(2), 85–94 (2009)
10. Thoma, G.R., Ford, G., Chung, M., Vasudevan, K., Antani, S.: Interactive Publications: Creation and Usage. In: *Electronic Imaging 2006*. pp. 607603–607603. International Society for Optics and Photonics (2006)
11. Vander Sande, M., Verborgh, R., Coppens, S., De Nies, T., Debevere, P., De Vocht, L., De Potter, P., Van Deursen, D., Mannens, E., Van de Walle, R.: Everything is Connected. In: Proceedings of the 11th International Semantic Web Conference (ISWC) (2012)