Towards Decentralized Annotations in Digital Books and on the Web

Tom De Nies[1], Ruben Verborgh[1], Miel Vander Sande[1], Ben De Meester[1],
Hajar Ghaem Sigarchian[1], Anastasia Dimou[1], Wesley De Neve[1,2],
Erik Mannens[1], and Rik Van de Walle[1]

[1] Ghent University – iMinds – Multimedia Lab, Belgium
{tom.denies,ruben.verborgh, miel.vandersande, hajar.ghaemsigarchian,
anastasia.dimou,wesley.deneve,erik.mannens,rik.vandewalle}@ugent.be
[2] KAIST – Image and Video Systems Lab, Republic of Korea

1 Introduction

In recent years, the concept of machine-interpretable annotations – for example using RDFa – has been gaining support in the Web community. Websites are increasingly adding these annotations to their content, in order to increase their discoverability and visibility to external agents and services.

This paper highlights two problems with current annotations and offers potential solutions. The first problem is that annotations largely remain *centralized*: the party responsible for publishing the content is also in control of the annotations. This limits the available annotation sources; for instance, note, comments, and links offered by third parties can only appear with the explicit approval of the content publisher. The second problem is that *digital books* have not undergone the same evolution yet, and the vast amount of useful information contained within them remains siloed and unaccessible for machines.

In the following section, we will discuss the need for decentralized annotations. Next, we highlight the importance for annotations in digital books.

2 Decentralizing annotation discovery

The Web's unidirectional nature often makes it difficult to find relevant annotations for specific content: including a link to a Web page or book from a certain comment, does not automatically add a back-link to that comment. So how can we find those comments and annotations that are relevant? One possibility to discover relationships between two resources in different directions are SPARQL endpoints; however, (i) their availability is currently problematic, and (ii) they only work on their local dataset, failing to provide federated querying. We therefore propose a more flexible and light-weight way of discovering annotations through *Linked Data Fragments* [7]. This is a strategy for offering Linked Data in reusable fragments; in this case, those fragments would be the comments and annotations for a specific Web page or (part of a) digital book.

Additionally, pages and books could be enhanced with third-party functionality, instead of only the action links provided by the information publisher. This becomes possible with *distributed affordance* [6].

Our vision on this matter resonates with the goals of the Crosscloud Project [1], which aims to move socially created data out of their silos, stimulating the movement of annotations across different applications on the Web. We believe that Linked Data Fragments and distributed affordance can play a part in this, by influencing the availability of data and actions related to specific content.

3 Creating machine-interpretable books

Within the Web community, it is well-known that a machine-interpretable version of content on the Web greatly improves the content's discoverability [5]. We strongly believe that the same is true for digital books, especially since the technologies used in the EPUB3 standard strongly relate to the Open Web standards. As argued by Hugh McGuire, a book can be viewed as an API to the content it contains [4]. The triples in an RDFa-annotated book can be flexibly queried, for instance, as Linked Data Fragments, as mentioned in the previous section.

To facilitate the addition of machine-interpretable annotations to an eBook, we should consider two aspects: the format of the book, and the format of the annotations. In our case, we argue that the use of HTML5/EPUB3 is the most feasible format for a machine-interpretable book, especially since this would also allow the use of the RDFa standard. Unfortunately, manually adding these RDFa annotations to eBooks – especially existing ones – is a very dull and labor-intensive process, and therefore infeasible in a realistic scenario. However, there are possible solutions to this.

First, the use of named-entity recognition would enable the automatic annotation of several types of information, including persons, locations, etc. While the recognized entities do require verification by a human annotator, they do exhibit sufficient accuracy to significantly reduce his/her workload [3]. Second, we incorporate HTML and RDFa annotations in the publishing workflow as early as possible. Machine-understandable semantics can already serve the automation, while the amount of data left to be converted to another format is drastically reduced. This allows querying and discovering the book structure (e.g., headings, footnotes), in addition to annotations in the actual content. In this context, we also investigate the possibilities in terms of template-driven HTML-to-RDF mappings, which could capture the semantics provided by HTML5. Therefore, we are developing novel authoring tools that support such a semantic, HTML5-based data model for books [2].

In addition to increased discoverability, machine-interpretable books offer much potential for drastically improving the user experience. For example, adding semantic annotations makes it possible to personalize the contents of a book with external, dynamic content. This content could originate from a repository controlled by the author or publisher, but also from other machine-interpretable resources on the open Web. Another example of this, is the addition of personalized actions, through distributed affordance [6]. If we elaborate on these concepts, one could even imagine books where the personalized contents, annotations and affordances become portable across all books in a user's personal collection.

4 Conclusion

As we have illustrated, there is great potential for machine-interpretable annotations in digital books in terms of discoverability and user experience. A key aspect herein is whether we will be able to provide these annotations in a decentralized way. We hope to further open the discussion on these topics, which we believe will create added incentive for adoption of the EPUB and Open Web Platform standards in the digital publishing domain.

References

[1] Barr, C., Hawke, S.: Introducing CrossCloud: A project to get your data out of silos (Jun 2013), *http://www.knightfoundation.org/blogs/knightblog/2013/6/25/introducing-crosscloud-project-get-your-data-out-silos/*

[2] 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)

[3] van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., Van de Walle, R.: Exploring entity recognition and disambiguation for cultural heritage collections. Literary and Linguistic Computing (2014), *http://freeyourmetadata.org/publications/named-entity-recognition.pdf*

[4] McGuire, H.: A publisher's job is to provide a good API for books: you can start with your index. The Indexer 31(1), 36–38 (2013)

[5] Rosati, A., Mayernik, M.: Facilitating data discovery by connecting related resources. Semantic Web Journal (2013)

[6] Verborgh, R., Hausenblas, M., Steiner, T., Mannens, E., Van de Walle, R.: Distributed affordance: An open-world assumption for hypermedia. In: Proceedings of the Fourth International Workshop on RESTful Design. pp. 1399–1406 (May 2013), *http://www2013.org/companion/p1399.pdf*

[7] Verborgh, R., Vander Sande, M., Colpaert, P., Coppens, S., Mannens, E., Van de Walle, R.: Web-scale querying through Linked Data Fragments. In: Proceedings of the 7th Workshop on Linked Data on the Web (Apr 2014)