Using semantic microformats for web archiving – an initial project conception

Márton Németh, National Széchényi Library, Budapest, Hungary

Abstract

In this short research paper, I would like to offer a basic overview about the definition of microdata and its way of use from library perspective in general. In the following I would like to describe some challenges in web archiving field from content display, research support and long term preservation perspectives. The main aim of this article to refer some completely new research and development activities in microdata support functions in web archiving context. Perhaps the International Internet Preservation Consortium (IIPC) Research Working group (IIPC RWG, 2019) could help to amalgamate some future steps in this field.

1. Microdata: the meeting point between document web and semantic web

Since the born of HTML standard some additional really basic metadata elements (for example by title statement) could be added to the header of the source code of a website (HTML 4.1 standard metadata elements, 2018). A robots.txt file has been attached to a website can help the archiving of a website by stating where the crawling robot can go and which segments are forbidden for them. These services have really attached to the document web where we can only take basic assumptions on the scope of an individual websites. Furthermore, these services are not helping us to retrieve info about the content of a website and we cannot link data from different websites together. The document web connects documents while links are unqualified. Semantic web (Berners Lee, Hendler& Lassila, 2001) is a next development step that is not focusing on individual document elements (websites) but focusing on the content and how web content segments can be identified, can be retrieved and can be catalogued as a part of a global information universe. Semantic web connects different datasets, instead of documents, with qualified links. In this way web-based resources can be handled together on a same platform with information from library catalogues or with any kind of standard information resources. Microdata (HTML Microdata W3C working draft, 2018) is appearing as a bridge among the document web and semantic web. It enables to make semantic statements as metadata embedded to the source code of a website and make its content available on the semantic web platform (Horváth, 2016). Microdata has defined in the environment of web standards and semantic web data models. RDFa (or Resource Description Framework in Attributes) is a W3C recommendation that adds a set of attribute-level extensions to HTML, XHTML and various XML-based document types for embedding rich metadata within Web documents. (XHTML+RDFa, 2015). This is the first major step. RDFa is a part of the RDF data-model a cornerstone of the semantic web. Mapping enables its use for embedding RDF subject-predicate-object expressions within XHTML documents. A really major advantage is that the predicate part of the RDF expressions can come from different vocabularies. In this way multiple vocabularies or schemas can be used on the same platform! One of the most important vocabularies is the schema.org. It has developed by Bing, Google and Yahoo!. Schema.org (Getting started with schema.org SA) is similar to a general thesaurus. The two most important value-added features of schema.org that it is suitable to describe any HTML pages of the Web!

Why microdata based on suitable vocabularies are important? The first important keyword is accuracy. HTML pages with schema.org markup are more understandable for search engines than other pages, on this way the search engines can give back more accurate search results. To index pages with standard microdata is so much easier for search engines and because of that these content resources are highly preferred by them. One of the first and biggest user of schema.org among library system providers is the Online Computer Library Center (OCLC). Every WorldCat description contains schema.org expressions (WorldCat Linked Data Vocabulary, SA). The result is that OCLC records are most likely are in the beginning of the search results of the general internet search engines like Bing, Google, etc. More and more open source library systems are following this practice by the same reason. Microdata however not just important because of indexing with high accuracy. As a part of the linked data universe several kind of semantic datasets can be linked with each other to offer such a large base of information retrieval that we have never had before. Furthermore, it is easy to realize that embedded information with microdata can offer relevant value added information to help long-term preservation of digital documents. In the following by this context we will focus on web archiving.

2. The potential ways of use of microdata related to web archiving

A Schema.org extension has published for the sub-domain for the bibliographic sector in 2015. (bib.schema.org. 1.0, 2015). This extended element set, together with core RDFa schema elements, has been integrated to several kinds of OPAC, discovery and repository system products (Koha, VuFind, Islandora, Dspace etc.). These content services can be identified and harvested by the help of permanent URLs, properly compiled sitemap and robots.txt settings, with RDFa and schema.org based marked up elements. Previously unharvested large bibliographic and full-text databases have become harvested on this way with qualified links containing bibliographic relations.

On the other hand, we still do not have a similar schema.org extension that can help to harvest any kind of other materials. OCLC has provided a basic metadata guideline for web archiving (Dooley & Bowers, 2018) that include schema.org mapping of the major proposed (mainly Dublin Core-based) elements. However, it offers only a really basic set of recommended metadata items. Schema.org or any other vocabulary do not include specific set of elements that can help to raise the effectiveness of web archiving. Create a new schema or extend an existing one requires some significant further research. In this short article a quick overview will be offered about in which context of web archiving challenges the microdata elements can appear.

2.1 Challenges with web archiving

Before the focus will turn on the use of microdata in web archiving context, in this chapter some major relevant challenges in web archiving are being overviewed related to potential use of microdata.

A major challenge is that robots cannot follow properly the whole structure of many websites because this structure has not specified clearly. There are further serious challenges of some old content formats that have become unsupported by the current browsers. Sometimes even it is not easy to determine what kind of original formats were used at all. We can only realize that some content is missing from the archived website, but the reasons can be really complex and not easily be determined because of the complexity of web as a medium. Another major challenge is the management of the large amount of data in big datasets. The content of these datasets must be a subject of research but make them researchable is a really complex task. The proper use of standard microdata schemas (vocabularies) can help to manage some of these above mentioned challenges. In the following challenge these various ways of use will be shortly presented.

2.2 Helping the setting of crawling by robots with microdata

Some types of microdata can help crawling robots in order to describe the complete structure of a website. A sitemap can be built-up in xml format. In the robots.txt file special set of commands can be inserted to help the handling of robot-specific microdata instructions. Robots.txt info content can be further developed, not just placing a separate file into the root directory, but some information can be inserted to the header to each page concerning the proper access to data. The additional information to robots. txt file (or a command can be put directly to the header of a page) are currently appearing as a set of non-standard extensions (Robots Exclusion Standard, Wikipedia). These are for example noindex metatag, noindex http response header and some directives just like allow, sitemap, host and universal match. These are not a part of the official standard, however, these must be added to the general description of it and must be available to all crawlers in all websites in all platforms. In this way for example a proper sitemap can be located. Permanent calendar apps, database links can be excluding from crawling to prevent to get lost in a crawler trap. The exclusion even would be managed on the header of a specific page with the standard form of noindex http response header. Furthermore, some additional extension of the robot’s exclusion standard would be really useful to determine the different kind of interfaces of a homepage (mobile, interface, barrier-free interface, crawler friendly interface). The robots would be set to crawl just a specific kind of interface or to crawl all, depending on the archiving policy.

By microdata the advertisement sections, pop-up warning, compulsory cookie management warning elements of a website can be marked up. In this way the content can be crawled with or without ads a pop-up content according to the archiving policy of an institution.

Specific instructions could be set for robots if a website owner just want a certain type of interface to get crawled. Different kind of language versions can be set and determine which language interfaces can be crawled. If the start page of a website only includes a simple query box, the robot cannot simply start to crawl this website. Robots can be redirected by a set of microdata to an interface where data records can be retrieved. In this way if the search and retrieval system cannot be preserved in its original form, data behind the interface can be crawled and preserved.

By setting these kind of information as microdata in a standardized way, the effectiveness of crawling with the recent software tools could be highly developed without arranging costly developing activities on the software developer side.

2.3 Helping the display of the archived material by microdata

A common challenge in web archiving that even if the crawling of materials is successful, the archived content cannot be displayed properly with the available software. Specific microdata elements can help to understand the structure and layout of the archived content. For example, in many cases special JavaScript or pdf add-ons are being used to retrieve embedded photos or documents. Pull-down menus show the structure of the content. By providing an alternative way of displaying photos, documents, offer a simple menu structure, and refer to them by microdata, the archived content can be totally retrieved.

2.4 Long term preservation support by microdata

Specific microdata can help to determine the different content formats in a website with the description of type, version and other important features. Short statements can be made about all types of software that are being used in a website. Problematic elements for long term preservation can be like embedded java applets, embedded flash applets and content can be identified with proper version and functions. These kind of specific details can be extremely important in the future, when conversion or emulation of content must be planned on a new preservation and service platform.

2.5 Research support of web archives with microdata

Supporting research use of web archives is an essential segment of a web archive service portfolio by any institution. Primarily serving semantic data can offer a major help to researchers for effective information retrieval from a huge set of resources and also offering connections among different kind of datasets in the linked data universe. The main tools in this sense are link relations as descriptive attributes that can be attached to a hyperlink in order to define the type of the link or the relationship among the source and destination resources. On the semantic web RDF typed links are fundamental in Linked Open Data datasets for identifying the relationship (predicate) type of RDF triples, contributing to the automatic processing ways of machine-readable statements of the Giant Global Graph on the Semantic Web. The typed links in RDF are expressed as the value of the rdf:type property, defining the relationship type using well-established controlled vocabulary terms or definitions from Linked Open Data datasets. (Microformats wiki, Link relations, 2011.) For example, personal names or geographical names in websites can be filtered and marked-up with various namespace schemas and ID's. Links to the semantic Wikipedia, or semantic library catalogs can also be added.

Another important aspect is to set proper data about the date of creation and date of last modification of a website. It is rather important for example to grant the credibility of archived material as a historical resource. Some current software that are being used to display the archived materials are representing various segments that crawled in different time periods by a joint display layout. Proper set of date of websites, webpages or even web page segments by microdata can enable to represent valid display of the archived documents from the same period.

Microdata can also support some research activities related to Digital Humanities. Digital philologists are using a large number of XML-based analyzing tools and procedures that can be also applied through proper semantic vocabularies as microdata. For example, a quotation in Latin or in Greek within an English text, several name forms of a same person or geographic place or different kind of embedded historical spelling forms could be identified in this way.

2.6 Establishing and applying microdata

When there is a focus on the use of microdata in web-archiving context a major question is that who can set vocabularies and managing the embedding of microdata. It is clear that the focus in this sense should be on the online content of public collections and large content provider platforms (like blog provider services or social media platforms). These stakeholders have the necessary resources to help establish vocabularies and implement their use. Only a widespread use of microdata can offer us effective benefits. In web archiving field IIPC can offer the major help to establish relevant vocabularies and build partnership with the major actors in the content development industry in order to implement microdata to their service portfolio.

Conclusion

This paper has focused on to show some recent major challenges and describe some new ways about the potential of using microdata in order to raise effectiveness of web archiving. The author's hope is that this paper can inspire to plan new research projects in this field and microdata can be effectively used in web archiving field in the future.

References

IIPC Research Working Group: http://netpreserve.org/about-us/working-groups/research-working-group/ retrieved in 23.09.2019

HTML 4.1. standard metadata elements: https://www.w3.org/TR/html401/struct/global.html#h-7.4.4 retrieved in 23.09.2019.

Lee, Berners, Tim, Hendler, James, Lassila, Ola. The Semantic Web. 2007. http://web.archive.org/web/20070713230811/http://www.sciam.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21 Retrieved in 23.09.2019

HTML Microdata W3C Working Draft. https://www.w3.org/TR/2018/WD-microdata-20180426/ Retrieved 23.09.2019.

Horváth, Ádám: RDFa – schema.org: unity of document and semantic web. Presentation on the Networkshop 2016. conference. https://conference.niif.hu/event/5/session/10/contribution/27/material/slides/0.ppt Retrieved 23.09.2019

XHTML+RDFa 1.1 – Third Edition. Support for RDFa via XHTML Modularization. W3C Recommendation 17.03. 2015. https://www.w3.org/TR/xhtml-rdfa/ Retrieved 23.09.2019.

Getting started with schema.org using microdata. https://schema.org/docs/gs.html Retrieved 23-09-2019.

OCLC WorldCat Linked Data Vocabulary. https://www.oclc.org/developer/develop/linked-data/worldcat-vocabulary.en.html Retrieved 23.09.2019.

Dooley, Jackie, and Kate Bowers. 2018. Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group. Dublin, OH: OCLC Research. https://doi.org/10.25333/C3005C. Retrieved 23.09.2019

Bib.schema.org-1.0 an initial release. 24.05.2015 https://www.w3.org/community/schemabibex/wiki/Bib.schema.org-1.0 Retrieved 23.09.2019

Robots Exclusion Standards. An article from Wikipedia. https://en.wikipedia.org/wiki/Robots_exclusion_standard Retrieved 23.09.2019

Link relation types. An article from Microformats Wiki. 20.09.2011. http://microformats.org/wiki/link-relation-types Retrieved 23.09.2019