# Metadata-Driven Hypertext Content Publishing and Styling

Xi Bai
BBC Publishing Services
Future Media
London W12 7TP, UK
xi.bai@bbc.co.uk

Armin Haller
Australian National University
CSIRO ICT Centre
Canberra ACT 2601, Australia
armin.haller@csiro.au

Ewan Klein, Dave Robertson
University of Edinburgh
Informatics Forum Edinburgh
EH8 9AB, UK
{ewan, dr}@inf.ed.ac.uk

## ABSTRACT

A growing number of approaches and tools have been utilised attempting at generating hypertext content with embedded metadata. However, little work has been carried out on finding a generic solution for publishing and styling Web pages with annotations derived from existing RDF data sets available in various formats. This paper proposes a metadata-driven publishing framework assisting publishers or webmasters in generating semantically-enriched content (HTML pages or snippets) by harnessing distributed RDF(a) documents or repositories with little human intervention. This framework also helps users to create and share so-called micro-themes, which is applicable to the above generated content for the purpose of page styling and also highly reusable thanks to the adopted semantic attribute selectors.

## Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services, Data sharing*

## General Terms

Algorithms, Design

## Keywords

Embedded metadata, micro-theme, linked data, RDF

## 1. INTRODUCTION

RDFa has started her life as a standard embeddable metadata format for XHTML2 since 2008 in the form of RDFa 1.0 and four years later, it evolved into RDFa 1.1 [1] which is able to work with more markup languages such as HTML5 and SVG. A number of applications dedicated to processing RDFa have been developed through leveraging the existing range of techniques for processing canonical RDF. However, publishing Web pages with RDFa still needs expertise and has become a non-trivial task for those publishers or webmasters lack of relevant knowledge. On the other hand, billions of RDF triples have been created but hidden either behind SPARQL endpoints or inside repositories and little attention has been paid to making use of existing triples as markups within the content publishing process. In this paper, a metadata-driven publishing and styling framework is proposed and it equips publishers with a lightweight tool, RDFa$^2$, which can help them in generating semantically-enriched hypertext content with annotations derived from existing distributed RDF(a) documents or repositories [3]. Without proper stylesheets, those generated content may look tedious and dreary and by no means meet modern Web design standards. Therefore, in this framework, a tool called Metastyle is also shifted and it can take the same set of vocabularies (involved in annotation generation via RDFa$^2$) as a seed and produce a skeleton for publishers to create a micro-theme dedicated to those vocabularies. Micro-themes have improved the readability and reusability of stylesheets by adopting semantic attribute selectors with URIs pointing to entities (or resources) and enabled easy-to-reuse themes.


*WWW'14 Companion*, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577344.

## 2. EMBEDDED METADATA GENERATION

We proposed an auto-templating algorithm for transforming RDF triples into HTML pages with annotations based on topic trees. Since the RDF data model relies on a graph which cannot be converted to one HTML DOM tree without duplicating re-entrant nodes, publishers are brought in to choose a specific node as the tree root of RDF statements. This type of nodes are named as *topic nodes*, each of which reflects a publisher's main topic of interest. The RDF document or repository from which a topic node is originated is called a *RDF context*, denoted by $\mathcal{C}$, and a set of RDF statements rooted in a topic node is called a *topic $\mathcal{C}$-tree*. There are two kinds of topic trees distinguished according to the topic-node position. Given a resource $r$, a context $\mathcal{C}$, and a RDF statement $(s, p, o)$, the *subject (topic) $\mathcal{C}$-tree* based on $r$ is defined as $\{(s, p, o) \in \mathcal{C} \mid s = r\}$, and similarly for the *object (topic) $\mathcal{C}$-tree* based on $r$. For instance, a subject topic tree from a FOAF document is depicted in Figure 1.

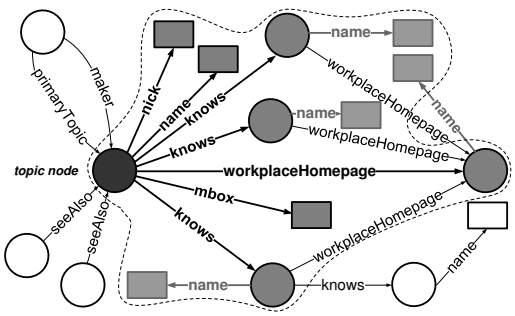


**Figure 1: *Subject (topic) $\mathcal{C}$-tree* in a FOAF document**

Figure 2 gives a screenshot on marked-up snippet generation from a Twitter profile already available in RDF with RDFa[2] via our framework. This snippet is ready to be inserted into a HTML `<body>` element or exported as a separate Web page. In each generated snippet, we also employ voiD[2] to include the relationships between embedded metadata and originated RDF contexts. Although an out-of-the-box free editor is available to publishers for customisation (as shown in Figure 2), the final Web pages may still look dreary due to the lack of proper stylesheets, and Section 3 is dedicated to tackling this issue with so-called micro-themes.

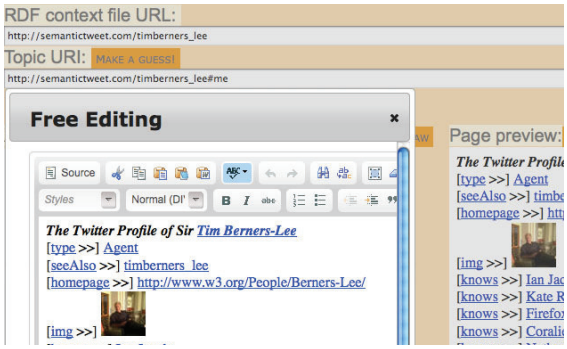


**Figure 2: Markup generation and customisation**

## 3. METADATA-DRIVEN THEMATISATION

With the help of RDFa[2], content publishers can get ready-to-publish marked-up snippets with little or no human intervention. As aforementioned, without any stylesheet being applied, those snippets will inevitably look tedious and by no mean meet the requirements of modern Web design. This is the place where Metastyle comes into play. In our framework, Metastyle can automatically generate stylesheet skeletons (in CSS or LESS) which can be applied to pages annotated with Microdata [4] or RDFa Lite [5]. More alternative data embeddable formats will be continually evaluated and integrated into our framework in the near future.

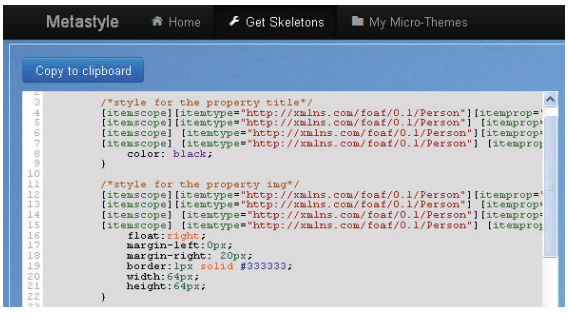


**Figure 3: Skeleton generation and customisation**

Figure 3 depicts the customisation from a publisher on an automatically generated stylesheet skeleton in CSS based on FOAF. Skeletons are also available in the LESS format which is more readable and compact compared with standard CSS. In our framework, the LESS variables for selectors follow the `prefix_class-property` naming convention (`prefix` stands for the prefix of a specific vocabulary namespace URI) in order to improve the readability and at the same time avoid potential selector clash especially when publishers import micro-themes into existing Web pages already with other stylesheets. Moreover, Microdata-based micro-themes and RDFa-based micro-themes can be transformed from one to the other thanks to the Metastyle *selector transformer*. The following extracted selector patterns form the foundation of this micro-theme transformation.

Microdata Item Selector:
`[itemscope][itemtype="`$TURI$`"]`
Microdata Property Selector:
`[itemscope][itemtype="`$TURI$`"][itemprop="`$PURI$`"]` |
`[itemscope][itemtype="`$TURI$`"]␣[itemprop="`$PURI$`"]`

RDFa Item Selector:
`[typeof="`$PRE:TNAME$`"]` | `[typeof="`$TNAME$`"]` | `[typeof="`$TURI$`"]`
RDFa Property Selector:
`[typeof="`$PRE:TNAME$`"][property="`$PRE:PNAME$`"]` |
`[typeof="`$PRE:TNAME$`"]␣[property="`$PRE:PNAME$`"]` |
`[typeof="`$TNAME$`"][property="`$PNAME$`"]` |
`[typeof="`$TNAME$`"]␣[property="`$PNAME$`"]` |
`[typeof="`$TURI$`"][property="`$PURI$`"]` |
`[typeof="`$TURI$`"]␣[property="`$PURI$`"]`

Here, $TURI$ stands for the URI of an entity type; $PURI$ and $PNAME$ stand for the URI and the local name of an entity property respectively; ␣ stands for whitespace characters; $PRE$ stands for the abbreviation of a namespace; $TNAME$ stands for the local name of an entity type.

## 4. CONCLUSION

Our framework provides the first one-stop solution (to the best of our knowledge) to hypertext content publishing driven by embedded metadata derived from existing RDF triples. RDFa[2][1] and Metastyle[2] have been shifted and integrated through this framework to assist publishers or webmasters in generating semantically-enriched Web pages, and also customising, sharing and transforming micro-themes applicable to those pages. 60 participants took part into our experiment on publishing exiting FOAF documents using this framework and 96.66% of them successfully got their profiles generated with annotations and indexed by Sindice[3]. Satisfactorily, 56 participants chose the first recommended topic nodes. Micro-themes currently applies the naming convention and the `!important` rule to CSS selectors to address the problem of name clashing and prioritise styling but if they were abused, stylesheets could become hard to maintain so a better way of fulfilling this will be investigated.

## 5. REFERENCES


[1] B. Adida, M. Birbeck, S. McCarron, and I. Herman. RDFa Core 1.1, W3C Recommendation. `http://www.w3.org/TR/rdfa-syntax/`, June 2012.

[2] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing linked datasets - on the design and usage of voiD, the 'vocabulary of interlinked datasets'. In *Proc. WWW Workshop on LDOW '09*, 2009.

[3] X. Bai. Addressing the rdfa publishing bottleneck. In *Proceedings of WWW'11*, pages 331–336. ACM, 2011.

[4] I. Hickson. HTML Microdata , W3C Working Draft. `http://www.w3.org/TR/microdata/`, October 2012.

[5] M. Sporny. RDFa Lite 1.1, W3C Recommendation. `http://www.w3.org/TR/rdfa-lite/`, June 2012.


[1] `http://demos.inf.ed.ac.uk:8836/rdfasquare/`
[2] `http://metastyle.cfapps.io/`
[3] `http://www.sindice.com/`