

A peer-reviewed version of this preprint was published in PeerJ on 2 October 2017.

[View the peer-reviewed version](https://peerj.com/articles/cs-132) (peerj.com/articles/cs-132), which is the preferred citable publication unless you specifically need to cite this preprint.

Peroni S, Osborne F, Di Iorio A, Nuzzolese AG, Poggi F, Vitali F, Motta E. 2017. Research Articles in Simplified HTML: a Web-first format for HTML-based scholarly articles. PeerJ Computer Science 3:e132
<https://doi.org/10.7717/peerj-cs.132>

Research Articles in Simplified HTML: a Web-first format for HTML-based scholarly articles

Silvio Peroni^{Corresp., 1}, **Francesco Osborne**², **Angelo Di Iorio**¹, **Andrea Giovanni Nuzzolese**³, **Francesco Poggi**¹, **Fabio Vitali**¹, **Enrico Motta**²

¹ Department of Computer Science and Engineering, University of Bologna, Bologna, Italy

² Knowledge Media Institute, Open University, Milton Keynes, United Kingdom

³ Institute of Cognitive Sciences and Technologies, Italian National Research Council, Rome, Italy

Corresponding Author: Silvio Peroni
Email address: silvio.peroni@unibo.it

Purpose: this paper introduces the Research Articles in Simplified HTML (or RASH), which is a Web-first format for writing HTML-based scholarly papers; it is accompanied by the RASH Framework , i.e. a set tools for interacting with RASH-based articles. The paper also presents an evaluation that involved authors and reviewers of RASH articles, submitted to the SAVE-SD 2015 and SAVE-SD 2016 workshops.

Design: RASH has been developed in order to: be easy to learn and use; share scholarly documents (and embedded semantic annotations) through the Web; support its adoption within the existing publishing workflow

Findings : the evaluation study confirmed that RASH can already be adopted in workshops, conferences and journals and can be quickly learnt by researchers who are familiar with HTML.

Research limitations: the evaluation study also highlighted some issues in the adoption of RASH, and in general of HTML formats, especially by less technical savvy users. Moreover, additional tools are needed, e.g. for enabling additional conversion from/to existing formats such as OpenXML.

Practical implications: RASH (and its Framework) is another step towards enabling the definition of formal representations of the meaning of the content of an article, facilitate its automatic discovery, enable its linking to semantically related articles, provide access to data within the article in actionable form, and allow integration of data between papers.

Social implications: RASH addresses the intrinsic needs related to the various users of a scholarly article: researchers (focussing on its content), readers (experiencing new ways for browsing it), citizen scientists (reusing available data formally defined within it through semantic annotations), publishers (using the advantages of new technologies as envisioned by the Semantic Publishing movement).

Value: RASH focuses strictly on writing the content of the paper (i.e., organisation of text + semantic annotations) and leaves all the issues about it validation, visualisation, conversion, and semantic data extraction to the various tools developed within its Framework.

1 Research Articles in Simplified HTML: 2 a Web-first format for 3 HTML-based scholarly articles

4 Silvio Peroni¹, Francesco Osborne², Angelo Di Iorio³, Andrea Giovanni
5 Nuzzolese⁴, Francesco Poggi⁵, Fabio Vitali⁶, and Enrico Motta⁷

6 ¹Digital And Semantic Publishing Laboratory, Department of Computer Science and
7 Engineering, University of Bologna, Bologna, Italy

8 ²Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom

9 ³Digital And Semantic Publishing Laboratory, Department of Computer Science and
10 Engineering, University of Bologna, Bologna, Italy

11 ⁴Semantic Technologies Laboratory, Institute of Cognitive Science and Technologies,
12 Italian National Research Council, Rome, Italy

13 ⁵Digital And Semantic Publishing Laboratory, Department of Computer Science and
14 Engineering, University of Bologna, Bologna, Italy

15 ⁶Digital And Semantic Publishing Laboratory, Department of Computer Science and
16 Engineering, University of Bologna, Bologna, Italy

17 ⁷Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom

18 Corresponding author:

19 Silvio Peroni¹

20 Email address: silvio.peroni@unibo.it

21 ABSTRACT

22 *Purpose:* this paper introduces the *Research Articles in Simplified HTML* (or *RASH*), which is a Web-
23 first format for writing HTML-based scholarly papers; it is accompanied by the *RASH Framework*, i.e.
24 a set tools for interacting with RASH-based articles. The paper also presents an evaluation that in-
25 volved authors and reviewers of RASH articles, submitted to the SAVE-SD 2015 and SAVE-SD 2016
26 workshops.

27 *Design:* RASH has been developed in order to: be easy to learn and use; share scholarly documents
28 (and embedded semantic annotations) through the Web; support its adoption within the existing publish-
29 ing workflow

30 *Findings:* the evaluation study confirmed that RASH can already be adopted in workshops, conferences
31 and journals and can be quickly learnt by researchers who are familiar with HTML.

32 *Research limitations:* the evaluation study also highlighted some issues in the adoption of RASH, and
33 in general of HTML formats, especially by less technical savvy users. Moreover, additional tools are
34 needed, e.g. for enabling additional conversion from/to existing formats such as OpenXML.

35 *Practical implications:* RASH (and its Framework) is another step towards enabling the definition of
36 formal representations of the meaning of the content of an article, facilitate its automatic discovery,
37 enable its linking to semantically related articles, provide access to data within the article in actionable
38 form, and allow integration of data between papers.

39 *Social implications:* RASH addresses the intrinsic needs related to the various users of a scholarly
40 article: researchers (focussing on its content), readers (experiencing new ways for browsing it), citizen
41 scientists (reusing available data formally defined within it through semantic annotations), publishers
42 (using the advantages of new technologies as envisioned by the Semantic Publishing movement).

43 *Value:* RASH focuses strictly on writing the content of the paper (i.e., organisation of text + semantic
44 annotations) and leaves all the issues about its validation, visualisation, conversion, and semantic data
45 extraction to the various tools developed within its Framework.

46 **RASH version:** <https://w3id.org/people/essepuntato/papers/rash-peerj2016.html>.

47 **INTRODUCTION**

48 In the last months of 2014, several posts within technical mailing lists of the Web¹ and Semantic Web²
49 community have discussed an evergreen topic in scholarly communication, i.e., how could authors of
50 research papers submit their works in HTML rather than, say, PDF, MS Word or LaTeX. Besides the
51 obvious justification of simplification and unification of data formats for drafting, submission and pub-
52 lication, an additional underlying rationale is that the adoption of HTML would ease the embedding of
53 semantic annotations, thus making a step towards the improvement of research communications thanks
54 to already existing W3C standards such as RDFa (Sporny, 2013), Turtle (Prud'hommeaux and Carothers,
55 2014) and JSON-LD (Sporny et al., 2014). This open complex and exciting scenarios that the Semantic
56 Publishing community has promised us in terms of increased discoverability, interactivity, openness and
57 usability of the scientific work (Bourne et al., 2011) (Shotton et al., 2009).

58 Nonetheless HTML is still primarily used as output format only: the authors write their papers in
59 LaTeX or MS Word and submit sources to the typesetters, who are responsible for producing the final
60 version, that eventually will be printed or read on the Web. Appropriate tools in the publishing toolchain
61 are used to convert papers among multiple formats.

62 The interest in *Web-first research papers* - that are natively designed, stored and transferred in HTML
63 - is increasing. Just to cite a few research efforts: Scholarly HTML³ defines a set of descriptive rules to
64 use a reduced amount of HTML for describing the metadata and content of scholarly articles; Dokieli⁴
65 is a Web application that allows authors to create HTML-based scholarly articles directly on the browser,
66 adding annotations and many other sophisticated features.

67 This paper introduces a novel approach towards the same goal: providing authors with a customized
68 version of HTML for Web-first papers. The format is called RASH, *Research Articles in Simplified*
69 *HTML*, and consists of a subset of 32 HTML elements only. This format is also accompanied by the
70 *RASH Framework*, i.e. a set of specifications and tools for RASH documents.

71 There are two key differences between RASH and other similar proposals. First of all, RASH adopts
72 a *simplified pattern-based data model*. The number of markup elements to be used by authors was
73 reduced down to the bare minimum, and the elements themselves were chosen in order to minimize
74 the cognitive effort of authors when writing documents. Secondly, RASH does not come with a full
75 authoring environment but is meant to be produced from MS Word, ODT and LaTeX sources. The basic
76 idea is to allow authors to keep using their own word processors, that are well known and allow them to
77 prepare articles in an easily and well-understood way, and to provide them with multi-format converters.
78 These converters are included in the RASH Framework, whose architecture is modular and extensible
79 for handling new formats in the future.

80 RASH is in fact designed to write the content of the papers only (i.e., organisation of text + semantic
81 annotations), handling all the issues about validation/presentation/conversion of RASH documents to the
82 various tools developed within its Framework. This is a well-established principle in scientific publishing:
83 *clear separation of concerns*. The authors must focus on organising the content and structure only, and
84 the format should not require authors to worry about how the content is presented on screen and print.
85 The publishers will then take care of creating the final formatting to best render the content in the style
86 of their publications.

87 This lead us to a further critical point valid for any HTML-based language to be used for scientific
88 writing: *good rendering and acceptance by the publishers*. Any new HTML-based format should be
89 beneficial for publishers as well. Of course publishers, conference and workshop organisers, would
90 like to manage new formats in the same way they already do for those formats they already support,
91 such as LaTeX. To this end, the new format should guarantee the possibility of developing tools for its
92 conversion and rendering into specific layouts, such as ACM ICPS⁵ and Springer LNCS⁶; RASH adopts
93 a *pragmatic approach* to solve this issue: while we are interested in a full-fledged native RASH authoring
94 environment, we implemented a set of converters, in the RASH Framework, that are easily integrable
95 (and were integrated) with existing publishing platforms.

¹<https://lists.w3.org/Archives/Public/public-lod/2014Nov/0003.html>

²<https://lists.w3.org/Archives/Public/public-lod/2014Oct/0058.html>

³<http://scholarlyhtml.org>

⁴<http://dokie.li>

⁵<http://www.acm.org/sigs/publications/proceedings-templates>

⁶<http://www.springer.com/computer/lncs?SGWID=0-164-6-793341-0>

96 The goal of this paper is in fact to describe the outcomes of some experimentations on the use of
97 RASH, so as to understand:

- 98 1. if it can be adopted as HTML-based submission format in academic venues (workshops, confer-
99 ences, journals);
- 100 2. if it is easy to learn and use;
- 101 3. if it can be used to add semantic annotations; in particular, what are the most widely adopted
102 vocabularies and how they were adopted in RASH papers;

103 The rest of the paper is structured as follows. In Section we introduce some of the most relevant
104 related works in the area, providing a functional comparison of the various works. In Section we
105 introduce the rationale for the creation of a new Web-first format for scholarly publication, discussing
106 the importance of *minimality*. In Section and Section we introduce the theoretical background of
107 RASH, and then provide an introduction of the language and the main tools included in its Framework.
108 In Section , as a case study, we discuss the use of RASH as one of the formats for submitting papers
109 to the SAVE-SD 2015⁷ and SAVE-SD 2016⁸ workshops. Finally, in Section we conclude the paper
110 sketching out some future developments.

111 RELATED WORKS

112 The growing interest in the publication of Web-first research papers have resulted in the release of some
113 interesting projects related to RASH. In the following subsections we discuss all the most important
114 contributions in this area by splitting them in two main categories: (i) HTML-based formats and (ii)
115 WYSIWYG editors for HTML documents.

116 Note that we do not discuss in detail some other efforts that have been recently done by means
117 of non-HTML languages, and are equally relevant for the community. ScholarlyMarkdown⁹ (Lin and
118 Beales, 2015), for instance, is a syntax to produce scholarly articles according to a Markdown¹⁰ input.
119 ShareLaTeX¹¹ is a Web-based real-time collaborative editor for LaTeX documents.

120 In Table 1 we briefly summarize the features and capabilities of the formats presented, in order to
121 highlight the main differences between them.

122 HTML-based formats

123 One of the first documented contributions that has proposed an HTML-based format for scholarly arti-
124 cles has been Scholarly HTML³⁰. It is not defined as a formal grammar, rather by a set of descriptive
125 rules which allows one to specify just a reduced amount of HTML tags for describing the metadata and
126 content of a scholarly article. It is the main intermediate format used in ContentMine³¹ for describing
127 the conversion of PDF content into HTML.

128 Along the same lines, PubCSS³² is a project which aims at pushing the use of HTML+CSS for writ-
129 ing scholarly articles. It does not define a formal grammar for the HTML element set to use. Rather it
130 provides some HTML templates according to four different CSS styles, which mimic four document La-
131 TeX stylesheets for Computer Science articles, i.e. ACM SIG Proceedings, ACM SIGCHI Proceedings,
132 ACM SIGCHI Extended Abstracts, and IEEE Conference Proceedings.

133 HTMLBooks³³ is an O'Reilly's specification for creating HTML documents (books, in particular)
134 by using a subset of all the (X)HTML5 elements. This is one of the first public works by a publisher
135 for pushing HTML-like publications, even if the status of its documentation (and, consequently, of its
136 schema) is still "unofficial".

137 Another project, that shares the same name of one of the previous ones, i.e. Scholarly HTML³⁴, is
138 a work by the science.ai³⁵ company that aims at providing a domain-specific data format based on open

⁷<http://cs.unibo.it/save-sd/2015/index.html>

⁸<http://cs.unibo.it/save-sd/2016/index.html>

⁹<http://scholarlymarkdown.com/>

¹⁰<http://daringfireball.net/projects/markdown/>

¹¹<https://www.sharelatex.com/>

³⁰<http://scholarlyhtml.org>

³¹<http://contentmine.org>

³²<https://github.com/thomaspark/pubcss/>

³³<https://github.com/oreillymedia/HTMLBook/>

³⁴<https://github.com/scienceai/scholarly.vernacular.io>

³⁵<http://science.ai>

Table 1. A comparison among existing HTML-oriented formats for scholarly papers according to seven distinct categories.

Format	Syntax	Documentation	Formal grammar	Semantic annotations	CSS for different formats	WYSIWYG editor	Conversion tools
RASH ¹²	HTML	Available online ¹³	RelaxNG ¹⁴	RDFa, RDF/XML, Turtle, JSON-LD	Web-based and Springer LNCS	OpenOffice	From: ODT To: LaTeX ACM ICPS and Springer LNCS
Scholarly HTML (2011) ¹⁵	HTML	Available online ¹⁶	None	RDFa	None	None	From: PDF (via Content-Mine - Norma ¹⁷)
PubCSS ¹⁸	HTML	Available online ¹⁹	Informal (via HTML templates)	None	ACM SIG Proceedings, ACM SIGCHI Proceedings, ACM SIGCHI Extended Abstracts, and IEEE Conference Proceedings	None	To: PDF (via browser interface)
HTMLBooks ²⁰	HTML	Available online ²¹	XML Schema ²²	None	CSS files for PDF printing and EPUB/MOBI-compatible device visualisations	None	None
Scholarly HTML (2015) ²³	HTML	Available online ²⁴	None	RDFa, JSON-LD	Web-based	Microsoft Word (as referenced online ²⁵)	From: DOCX
Scholarly HTML (2016) ²⁶	HTML	Available online ²⁷	None	RDFa, JSON-LD	Web-based	None	None
dokieli format	HTML	None	Informal (via HTML templates)	RDFa, Turtle, JSON-LD, TRiG	Web-based, Springer LNCS, ACM ICPS	dokieli ²⁸	To: PDF (via browser interface)
Fiduswriter format	HTML	None	None	None	Web-based	Fiduswriter ²⁹	To: HTML, EPUB, LaTeX

standards (among which HTML5) for enabling “the interoperable exchange of scholarly articles in a manner that is compatible with off-the-shelf browsers” (Berjon and Ballesteros, 2015). While the format it is not defined by any particular formal grammar, it has a well-described documentation (Berjon and Ballesteros, 2015) that teaches how to produce scholarly documents by using a quite large set of HTML tags, accompanied by schema.org³⁶ annotations for describing specific structural roles of documents as well as basic metadata of the paper. In services made available by the company would enable also the conversion from Microsoft Word document into such ScholarlyHTML format.

One of the authors of the previous work is also the chair of a W3C community group called “Scholarly HTML”³⁷ which aims at developing a HTML vernacular³⁸ for the creation of a Web-first format

³⁶<http://schema.org>

³⁷<https://www.w3.org/community/scholarlyhtml/>

³⁸<https://github.com/w3c/scholarly-html>

148 for scholarly articles. It involves several people from all the aforementioned specifications (including
149 RASH), and the group work should result in the release of a community-proposed interchange HTML
150 format. As of August 10, 2016, the online documentation³⁹ is mainly a fork of the Scholarly HTML
151 specification proposed by science.ai discussed above.

152 HTML-oriented WYSIWYG editors

153 One of the most important and recent proposals which is compliant with the principles introduced as part
154 of the *Linked Research*⁴⁰ project⁴¹ (Capadisli et al., 2015) is *dokieli*⁴². Dokieli is Web applications (still
155 under-development) that allows one to create HTML-based scholarly articles directly on the browser,
156 and implement several features among which annotations (in RDF), a notification system, and other fea-
157 tures. The application makes available also available some HTML templates and a series of widgets for
158 navigating, visualising (in different formats) and printing research documents easily by using common
159 browsers.

160 Fidus Writer⁴³ is another Web-based application for creating HTML scholarly documents by means
161 of a wordprocessor-like interface. While the particular format used is not explicitly specified, it allows
162 one to convert the HTML documents created within the application in two different formats, i.e. EPUB
163 and LaTeX (alongside with HTML).

164 WHICH “WEB-FIRST” FORMAT FOR RESEARCH ARTICLES?

165 The term “Web-first” format indicates the possibility of using HTML as primary language to write, store
166 and transfer research articles, and not only to make these articles available on the Web. Some questions
167 naturally arise in this context: shall we use the full HTML? If we limit to a subset, which elements should
168 we consider? Shall we demand specific rules for using the language?

169 Some works, e.g. (Capadisli et al., 2015), suggest not to force any particular HTML structure for
170 research papers. This choice would allow authors to use whatever HTML structure they want for writing
171 papers, and would reduce (even, eliminate) the fear for the *template bottleneck*, i.e., the fact that users
172 will not adopt a particular language because they would be obliged to follow specific rules. However,
173 leaving the user (i.e., the author) the freedom of using, potentially, the whole HTML specification may
174 affect, in some way, the whole writing and publishing process of articles.

175 First of all, the author of a paper is free to use any possible kinds of HTML *linearisations* for her
176 scholarly text, e.g.: using elements `div` instead of elements `section`, using elements `table` for
177 presentational behaviour and not for presenting tabular data, and the like. This freedom could, thus,
178 result in two main kinds of issues:

- 179 • *visualisation bottleneck* – it may affect the correct use of existing, well-developed and pretty stan-
180 dard CSSs (e.g., Capadisli’s CSSs developed for Dokieli⁴⁴) for both screen and print media, result-
181 ing in writing new code for handling paper visualisation correctly;
- 182 • *less focus on the research content* – the fact that a certain paper is not visualised in a browser very
183 well (or, worse, in a way that is not the one the author expects) could bring the author to work on
184 the presentation of the text, rather than on focussing on the actual research context of the text.

185 Another point against the use of any HTML syntax for writing papers concerns the possibility of
186 enabling an easy way for sharing the paper to others (e.g., co-authors) who, potentially, may not use
187 HTML in the same way. If all the co-authors of a paper are able to use the full HTML, they may not
188 understand other users’ specific uses of some HTML tags — “why did she use the elements `section`
189 instead of `div`?”; “what is this freaky use of elements `table`?”. Hence, the advantages of using a
190 common HTML format is quite evident: only one syntax and only one possible semantics.

191 There is a further issue worth mentioning. Having a simple and acceptable format would facili-
192 tate conversions from/into other complex ones (e.g., ODT (JTC1/SC34 WG 6., JTC1/SC34 WG 6.),

³⁹<https://w3c.github.io/scholarly-html/>

⁴⁰<http://linkedresearch.org>

⁴¹The main aim of the LinkedResearch project is to propose principles for enabling researchers to share and reuse research knowledge by means of existing Web and Semantic Web technologies towards a future world where researchers can publish and consume human-friendly and machine-readable (e.g., by using RDFa (Sporny, 2013)) scholarly documents.

⁴²<http://dokie.li>

⁴³<https://www.fiduswriter.org/>

⁴⁴<http://dokieli.io>

193 OOXML (JTC1/SC34 WG 4., JTC1/SC34 WG 4.), DocBook (Walsh, 2009), JATS (National Informa-
 194 tion Standards Organizatio, 2012)), thus enabling authors to use their own text editors or word-processors
 195 to modify the articles. The conversion is instead much more complex, error-prone and imprecise on the
 196 full HTML.

197 To complicate an already complex scenario is the *necessary* involvement of publishers. Leaving
 198 the authors of using their own HTML format could be also counterproductive from a publisher's per-
 199 spective, in particular when we speak about the possibility of adopting such HTML formats for regular
 200 conference/journal camera ready submissions. From a recent discussion on the Force11 mailing list⁴⁵,
 201 it appears clear that publishers are willing to adopt HTML for submissions *if and only if* it is a clear
 202 community need. It means that they will include HTML formats in the publishing workflow only once a
 203 number of conference organisers decide to deliver HTML as camera ready versions of accepted papers⁴⁶.
 204 However, using one clear Web-first format, rather than a plethora of possible variations allowed by the
 205 full HTML schema, would certainly decrease the effort of publishers for including HTML within the
 206 publishing workflow. This inclusion could be additionally supported by the community itself if it would
 207 be made available a series of services (e.g., converters, enhancers, visualisers) for facilitating the use of
 208 such Web-first format within the existing publishing environment.

209 Last but not least, using a controlled subset of HTML is more appropriate for *Semantic Publishing*
 210 applications (Shotton et al., 2009) (Peroni, 2014). The development of scripts and applications to extract,
 211 for instance, RDF statements directly from the markup structure of the text is a sort of nightmare if dif-
 212 ferent authors use HTML in different manners. For instance, what happen if we would like to extract the
 213 rhetorical organisation of a scientific paper according to the Document Component Ontology (DoCO)⁴⁹
 214 (Constantin et al., 2016) from two HTML documents that use HTML tags in different ways? Is an HTML
 215 element `table` an actual table (containing tabular data)? What are the tags identifying sections? These
 216 analyses are all easier on a controlled subset of HTML.

217 WRITING SCHOLARLY ARTICLES IN HTML WITH RASH

218 The subset of HTML we propose in RASH is strictly compliant to a *patterns theory*, we developed in
 219 the past years. In this section we briefly introduce these theoretical foundations and then we go into the
 220 details of RASH.

221 Theoretical foundations: structural patterns

222 While we have plenty of tools and languages for creating new markup languages (e.g. RelaxNG (Clark
 223 and Makoto, 2001) and XMLSchema (Gao et al., 2012)), they usually do not provide any particular
 224 guideline for fostering the development robust and well-shaped document languages. In order to fill that
 225 gap, in the last decade we have experimented the use of a *theory of structural patterns* for markup docu-
 226 ments (Di Iorio et al., 2014), that then has been already applied in a bunch of national and international
 227 standards, e.g. OASIS LegalDocumentML⁵⁰⁵¹, which is a legal document standard for the specifica-
 228 tion of parliamentary, legislative and judicial documents, for their interchange between institutions in
 229 different countries.

230 The basic idea behind this theory is that each element of a markup language should comply with one
 231 and only one structural pattern, depending on the fact that the element:

- 232 • can or cannot contain text (+*t* in the first case, -*t* otherwise);
- 233 • can or cannot contain other elements (+*s* in the first case, -*s* otherwise);
- 234 • is contained by another element that can or cannot contain text (+*T* in the first case, -*T* otherwise).

235 By combining all these possible values – i.e. $\pm t$, $\pm s$, and $\pm T$ – we basically obtain eight core
 236 structural patterns, namely (accompanied by a plausible exemplar within the HTML elements):

⁴⁵<https://groups.google.com/forum/#!topic/forcnet/g4BNAO0MjMM>

⁴⁶Note that accepting HTML as format for submissions in conferences/workshops is a totally different issue, since this choice is normally taken by the organisers. For instance, see the SAVE-SD 2015 call for papers⁴⁷ and the various editions of SePublica⁴⁸.

⁴⁹<http://purl.org/spar/doco>

⁵⁰<https://www.oasis-open.org/committees/legaldocml/>

⁵¹OASIS LegalDocumentML is the standardisation of AkomaNtoso⁵², which is a set of simple technology-neutral electronic representations in XML format of parliamentary, legislative and judiciary documents, and has been already adopted by several parliaments in European Union, Africa, and South America.

- 237 1. inline [+t+s+T], e.g. the element `em`;
- 238 2. block [+t+s-T], e.g. the element `p`;
- 239 3. popup [-t+s+T], e.g. the element `aside`;
- 240 4. container [-t+s-T], e.g. the element `section`;
- 241 5. atom [+t-s+T], e.g. the element `abbr`;
- 242 6. field [+t-s-T], e.g. the element `title`;
- 243 7. milestone [-t-s+T], e.g. the element `img`;
- 244 8. meta [-t-s-T], e.g. the element `link`.

245 Instead of defining a large number of complex and diversified structures, the idea is that a small
246 number of structural patterns are sufficient to express what most users need. Therefore, the two main
247 aspects related to such patterns are:

- 248 • orthogonality – each pattern has a specific goal and fits a specific context. It makes it possible to as-
249 sociate a single pattern to each of the most common situations in document design. Conversely, for
250 every situation a designer encounters in the creation of a new markup language, the corresponding
251 pattern is immediately selectable and applicable;
- 252 • assemblability – each pattern can be used only in some contexts within other patterns. This strict-
253 ness provides expressiveness and non-ambiguity in the patterns. By limiting the possible choices,
254 patterns prevent the creation of uncontrolled and misleading content structures.

255 Such patterns allow authors to create unambiguous, manageable and well-structured markup lan-
256 guages and, consequently, documents, and basically allow to increase the reusability (e.g., inclusion,
257 conversion, etc.) among different languages. Also, thanks to the regularity they provide, it is possible
258 to perform easily complex operations on pattern-based documents even when knowing very little about
259 their vocabulary (automatic visualisation of document, inferences on the document structure, etc.). Thus
260 designers can implement more reliable and efficient tools, can make hypothesis regarding the meanings
261 of document fragments, can identify singularities and can study global properties of sets of documents.

262 We applied these guidelines for restricting HTML – which is not pattern-based at all, since it allows
263 the creation of arbitrary and, sometimes, quite ambiguous structures – and define RASH, so as to select
264 a good subset that is enough expressive to capture the typical components of a scholarly article and that
265 is also well-designed, easy to reuse and robust at the same time.

266 **RASH: Research Article in Simplified HTML**

267 The *Research Articles in Simplified HTML (RASH)* format is a markup language that restricts the use of
268 HTML⁵³ elements to only 32 elements for writing academic research articles. It allows authors to use
269 RDFa⁵⁴ annotations (Sporny, 2013) within any element of the language⁵⁵. In addition to RDFa, RASH
270 makes available another way to add RDF statements to the document, i.e., the use of an element `script`
271 (with the attribute `type` set to “application/rdf+xml”, “text/turtle” or to “application/ld+json”) within
272 the element `head` for adding plain RDF/XML (Gandon and Schreiber, 2014), Turtle (Prud’hommeaux
273 and Carothers, 2014) or JSON-LD content (Sporny et al., 2014). In addition, RASH strictly follows the
274 Digital Publishing WAI-ARIA Module 1.0 (Garrish et al., 2016) (which is currently a working draft) for
275 expressing structural semantics on various markup elements used.

276 Any RASH documents begins as a simple (X)HTML5 document⁵⁶, by specifying the generic HTML
277 DOCTYPE followed by the document element `html` with the usual namespace (i.e. “http://www.w3.org/
278 1999/xhtml”) and with additional (and mandatory) prefix declarations through the attribute `prefix`⁵⁷.
279 The element `html` contains the element `head` for defining metadata of the document according to

⁵³<http://www.w3.org/TR/html5/>

⁵⁴<http://www.w3.org/TR/rdfa-syntax/>

⁵⁵Technically speaking, this is a meta-article, since it has been actually written by using RASH itself as markup language. Thus it is possible to easily access the HTML code of this article to understand how the various elements are rendered by the browser.

⁵⁶Please refer to the official RASH documentation, available at <http://cs.unibo.it/save-sd/rash>, for a complete introduction of all the elements and attributes that can be used in RASH documents.

⁵⁷The following prefixes are always mandatory in any RASH document:

- `schema`: <http://schema.org/>
- `prism`: <http://prismstandard.org/namespaces/basic/2.0/>

280 the DCTERMS⁵⁸ and PRISM⁵⁹ standards, and the element `body` for including the whole content of
 281 the document. On the one hand, the element `head` of a RASH document must/should include some
 282 information about the paper, i.e., the paper title (element `title`), at least one author and other related
 283 information (i.e., affiliations, keywords, categories, by using the elements `meta` and `link`). On the
 284 other hand, the element `body` mainly contains textual elements (e.g., paragraphs, emphases, links, and
 285 quotations) for describing the content of the paper, and other structural elements (e.g., abstract, sections,
 286 references, and footnotes) used to organise the paper in appropriate blocks and to present specific
 287 complex structures (e.g., figures, formulas, and tables).

288 In the following subsection we provide a quick discussion about pattern-usage in RASH, and we
 289 introduce the tools used for developing its grammar.

290 **Development and patterns**

291 As already mentioned, the development of RASH started from the whole HTML5 grammar, and pro-
 292 ceeded by removing and restricting the particular use of HTML elements, so as to be enough expressive
 293 for representing the structures of scholarly papers and to have the language totally compliant with the
 294 theory on *structural patterns* for XML documents (Di Iorio et al., 2014) introduced in Section

295 As already introduced, the systematic use of these structural patterns is an added value in all stages
 296 of the documents' lifecycle: they can be guidelines for creating well-engineered documents and vocabu-
 297 laries, rules to extract structural components from legacy documents, indicators to study to what extent
 298 documents share design principles and community guidelines. All these characteristics have allowed
 299 us to simplify, at least to some extent, the handling of all the requirements introduced in Section and
 Section in RASH. Table 2 shows what is the current pattern assignment for each element in RASH.

Table 2. The use of structural patterns in RASH.

Pattern	RASH element
inline	a, code, em, math, q, span, strong, sub, sup, svg
block	figcaption, h1, p, pre, th
popup	none
container	blockquote, body, figure, head, html, li, ol, section, table, td, tr, ul
atom	none
field	script, title
milestone	img
meta	link, meta

300 As shown, we do not use some of the patterns presented in Section , i.e. *atom* and *popup*. The ele-
 301 ments compliant with the former pattern are usually defined for describing textual content in discursive
 302 blocks (e.g. paragraphs) but prohibit to contain additional elements. Considering the context of schol-
 303 arly writings, this is a quite odd situation, since usually any element used for emphases, links, and other
 304 in-sentence elements can always contain additional ones.

305 A different discourse can be done for the pattern *popup*, which is used for any structure that, while
 306 still not allowing text content inside itself, is nonetheless found in elements with a mixed content context
 307 [t+s+], and it is meant to represent complex substructures that interrupt but do not break the main flow
 308 of the text, such as footnotes (Di Iorio et al., 2014). In particular, in developing RASH, we discussed
 309 which of the following two possible approaches for defining footnotes had to be more reasonable for our
 310 needs.

311 The first, is a *container*-based behaviour, also suggested by JATS (National Information Standards
 312 Organizatio, 2012) by means of the element `fn-group`, that allows one to specify footnotes (through
 313 the element `ft`) by using a tag that is totally separated from the main text from which it is referenced
 314 (usually through XML attributes), as shown in the following excerpt:

```
316 <-- A paragraph referring to a footnote -->
317 <p>
```

⁵⁸<http://dublincore.org/documents/dcmi-terms/>

⁵⁹<http://www.prismstandard.org/>

```

318     In this paragraph there is an explicit reference to the
319     second footnote<xref rid="n2"></xref>.
320 </p>
321
322 <-- The group containing all the footnotes -->
323 <fn-group>
324   <fn id="n1">
325     <p>This is a paragraph within a footnote.</p>
326   </fn>
327   <fn id="n2">
328     <p>This is a paragraph in another footnote.</p>
329   </fn>
330     All the footnotes are contained in a group, so as
331     to collect them together.
332   </p>
333 </fn-group>
334 ...
335 </fn-group>

```

336 The alternative is, in fact, a *popup*-based behaviour, used as default in LaTeX (by using the marker
337 `\footnote{}`) and even possible in JATS (which is a very permissive language by design), where a
338 paragraph can be abruptly interrupted by one or more paragraphs specified in a footnote, as shown in the
339 following excerpt:

```

340 <-- A paragraph containing a footnote -->
341 <p>
342   In this paragraph the footnote <fn id="n3"><p>That is
343   what we call popup-based behaviour!.</p></fn> has been
344   defined directly within it.
345 </p>

```

346 In RASH, we considered the latter approach a bit confusing, since it actually decreases the readability
347 of the HTML source where footnotes are needed, and thus decided to adopt a solution similar to the JATS
348 `fn-group` element, introduced as follows:

```

349 <-- A paragraph referring to a footnote -->
350 <p>
351   In this paragraph there is an explicit reference to the
352   second footnote<a href="#fn2"></a>.
353 </p>
354
355 <-- The group containing all the footnotes -->
356 <section role="doc-footnotes">
357   <section role="doc-footnote" id="fn1">
358     <p>This is the text of a footnote.</p>
359   </section>
360   <section role="doc-footnote" id="fn2">
361     <p>This is the text of another footnote.</p>
362   </section>
363   ...
364 </section>

```

365 **Grammar and peculiarities**

366 The formal grammar of RASH⁶⁰ (current version: 0.5) has been developed by means of RelaxNG (Clark
367 and Makoto, 2001), which is a simple, easy to learn, and powerful schema language for XML. The gram-
368 mar has been logically organised in four distinct logical blocks of syntactic rules, defining respectively
369 elements, attributes, content models⁶¹ for the elements and their related attribute lists, as summarised in
370 the following excerpt:

```

371 ...
372 <define name="p">
373   <element name="p">
374     <ref name="attributes_html_element_no_role" />
375     <ref name="cm_inline" />
376   </element>
377 </define>
378 ...
379 <define name="aClass">
380   <attribute name="class">
381     <data type="NMTOKENS" />
382   </attribute>
383 </define>

```

⁶⁰<https://raw.githubusercontent.com/essepuntato/rash/master/grammar/rash.rng>

⁶¹The *content model* of an element is the particular organisation of its content in terms of text, attributes and elements that it can contain.

```

384 ...
385 <define name="cm_inline">
386   <zeroOrMore>
387     <choice>
388       <text />
389       <ref name="a" />
390       <ref name="aRef" />
391       <ref name="img" />
392       <ref name="svg" />
393       <ref name="math" />
394       <ref name="img_math" />
395       <ref name="span_latex" />
396       <ref name="span" />
397       <ref name="code" />
398       <ref name="sub" />
399       <ref name="sup" />
400       <ref name="em" />
401       <ref name="strong" />
402       <ref name="q" />
403     </choice>
404   </zeroOrMore>
405 </define>
406 ...
407 <define name="attributes_html_element_no_role">
408   <ref name="attributes_html_generic" />
409   <optional>
410     <ref name="aClass" />
411   </optional>
412   <ref name="attributes_rdfa" />
413 </define>
414 ...

```

Starting from the latest versions of the language, there has been a clear shift so as to use more HTML5 semantic elements, despite the fact they are not back-compatible with possible (and more generic) alternatives in HTML4. In particular, the elements `section`, `figure`, and `figcaption` have been adopted so as to clearly refer to paper sections and boxes with tables, figures, listings and formulas, accompanied by a particular caption.

While this choice has fostered the readability of the source, the use of these HTML5 elements was not enough to guarantee a proper semantics and accessibility to the RASH source. Thus, in order to improve the user experience in terms of accessibility of such HTML-based papers, RASH reuses some items from the W3C Accessible Rich Internet Applications 1.1 (Diggs et al., 2015), and also exploits several roles introduced in the Digital Publishing WAI-ARIA Module 1.0 (Garrish et al., 2016), which allows “digital publisher to apply the structural semantics they need to drive the authoring process while getting accessibility for free”⁶². The use of such semantics is implemented by means of the attribute `@role`⁶³, that can be used on certain RASH elements, e.g. `sections`, and it is very useful for specifying a clear structural semantics where it is not formally defined. For instance, all the references are organised in a list within a special section defined by using the element `section` with the attribute `@role` set to “doc-bibliography”. This special section contains one list with a bibliographic reference for each list item (i.e., the element `li` accompanied by the attribute `@id` for referencing to it within the text and the attribute `@role` set to “doc-biblioentry”), as shown in the following excerpt:

```

433 <section role="doc-bibliography">
434   <h1>References</h1>
435   <ol>
436     <li id="Per2014" role="doc-biblioentry">
437       <p>Write here the reference entry.</p>
438     </li>
439     ...
440   </ol>
441 </section>

```

Formulas have been taken in particular consideration, since different ways are possible so as to implement them. The standard specification for representing mathematics on the Web is MathML (Carlisle et al., 2014). Even if MathML is the best accessible way for writing mathematical formulas, the organisation of the elements for defining even a quite simple formula is quite verbose and this is a reasonable obstacle to its direct adoption, as shown in the following excerpt for describing the formula r^2 :

```

442 <math xmlns="http://www.w3.org/1998/Math/MathML">

```

⁶²<https://lists.w3.org/Archives/Public/public-dpub-aria/2016Feb/0032.html>

⁶³In the paper, for the sake of readability, we use the prefix “@” when we name attributes (e.g. the attribute named “role” is introduced as `@role`), while we just name elements with their name (e.g. `section`).

```

448     <mi></mi>
449     <mo><!-- &InvisibleTimes; --></mo>
450     <msup>
451         <mi>r</mi>
452         <mn>2</mn>
453     </msup>
454 </math>

```

So as to help the creator of RASH documents in dealing with formulas, RASH adds other two ways for writing formulas in addition to MathML. The first one is to use an image (element `img`), which is a very simple way to include maths in a paper. On the other hand, it is not accessible at all since the various elements of the formula are not marked-up properly so as to distinguish them. Another option would be to use LaTeX, which is one of the most common ways to write formulas in many scientific papers. Both the options are specifiable in RASH by using either the element `img` or the element `span` respectively, accompanied by the attribute `@role` set to “math”, as shown in the following excerpt:

```

462 <!-- Specifying a formula through the element 'img' -->
463 
464
465 <!-- Specifying a formula in LaTeX through the element 'span' -->
466 <span role="math">\pi r^2</span>

```

The rendering of any LaTeX formula and the multi-browser support for MathML is implemented by using MathJax⁶⁴, which is a Javascript display engine for mathematics that works in all browsers. Of course, it is necessary to explicitly import it in the element `head` if any rendering of formulas is actually needed, as shown as follows:

```

471 <!-- MathJax for multi-browser support of LaTeX formulas and MathML -->
472 <script src="https://cdn.mathjax.org/mathjax/latest/MathJax.js?config=TeX-AMS-MML_HTMLorMML"> </
473 script>

```

RASH has been developed in order to allow anyone to add RDFa annotations (Sporny, 2013) to any element of the document. For instance, this paragraph contains the following RDF statements (in Turtle (Prud’hommeaux and Carothers, 2014)):

```

477 @prefix cito: <http://purl.org/spar/cito/> .
478 <> cito:credits <http://www.w3.org/TR/rdfa-syntax/> .

```

That was implemented by using specific RDFa attributes (`@property` and `@resource`, in this case) within the paragraph content, while the prefixes were defined in the element `html`, as shown in the following excerpt:

```

482 <html prefix="cito: http://purl.org/spar/cito/">
483     ...
484     <p>
485         RASH has been developed in order to allow anyone to add
486         <span
487             property="cito:credits"
488             resource="http://www.w3.org/TR/rdfa-syntax/">RDFa</span>
489         annotations to any element of the document.
490     </p>
491     ...
492 </html>

```

In addition to RDFa, RASH makes available another way to inject RDF statements (Cyganiak et al., 2014) to the document, by means of an element `script` (within the element `head`):

- with the attribute `type` set to “text/turtle” for adding plain Turtle content (Prud’hommeaux and Carothers, 2014);
- with the attribute `type` set to “application/ld+json” for adding plain JSON-LD content (Sporny et al., 2014);
- with the attribute `type` set to “application/rdf+xml” for adding plain RDF/XML content (Gandon and Schreiber, 2014).

An example of use of `script` for Turtle and JSON-LD statements is introduced in the following excerpt:

```

503 <script type="text/turtle">
504     @prefix pro: <http://purl.org/spar/pro/> .
505     @prefix foaf: <http://xmlns.com/foaf/0.1/> .

```

⁶⁴<https://www.mathjax.org/>

```

506     @prefix sd: <https://w3id.org/scholarlydata/person/> .
507     sd:silvio-peroni a foaf:Person ;
508         foaf:givenName "Silvio" ;
509         foaf:familyName "Peroni" ;
510         foaf:homepage <http://www.essepuntato.it> ;
511     pro:holdsRoleInTime [
512         a pro:RoleInTime ;
513         pro:withRole pro:author ;
514         pro:relatesToDocument <>
515     ] .
516 </script>
517
518 <script type="application/ld+json">
519     {
520         "@context":
521         {
522             "nick": "http://xmlns.com/foaf/0.1/nick",
523             "sd": "https://w3id.org/scholarlydata/person/"
524         },
525         "@id": "sd:silvio-peroni",
526         "nick": ["S.", "essepuntato"]
527     }
528 </script>

```

529 It is worth noticing that, excepting three properties from schema.org⁶⁵ for defining author's meta-
530 data (see Section 2 of the RASH documentation⁶⁶ for additional details), RASH does not constrain any
531 particular vocabulary for introducing RDF statements. For instance, in this document (in particular, in
532 its RASH version⁶⁷) we mainly use CiTO (Peroni and Shotton, 2012) and other SPAR Ontologies
533 (Peroni, 2014) for creating citation statement about the paper itself, but alternative and/or complementary
534 vocabularies are freely usable as well.

535 THE RASH FRAMEWORK

536 One of the issues we had to face, and in general anyone has to face when proposing a new markup
537 language, was to provide tools for writing papers in RASH. It is undeniable that:

- 538 • not all the potential authors are able (or willing) to write scholarly articles in HTML, even consid-
539 ering those people within the Web community;
- 540 • not all the potential authors are able (or willing) to add additional semantic annotations, even
541 considering the Semantic Web community.

542 The authorial activity of writing an article by using RASH, but also any other new Web-first format,
543 *must* be supported by appropriate interfaces to reach a broad adoption.

544 One possible solution could have been to implement a native HTML authoring environment, so as
545 authors do not have to deal with the new language directly. Apart from the technical difficulties in
546 creating such environment and in making it acceptable/accepted by the final users, there is another issue:
547 all co-authors are required to stick to the same tool. We believe that a more liberal approach, that allows
548 each author to keep using her/his preferred tools, even off-line, is more practical.

549 This is the idea behind the RASH Framework⁶⁸: a set of specifications and writing/conversion/extrac-
550 tion tools for writing articles in RASH. In this section we give a brief description of all the tools we have
551 developed in the framework. All the software components are distributed under an ISC License⁶⁹, while
552 the other components are distributed under a Creative Commons Attribution 4.0 International License⁷⁰.
553 A summary of the whole framework is introduced in Fig. 1.

554 Validating RASH documents

555 RASH has been developed as a RelaxNG grammar (Clark and Makoto, 2001), i.e., a well-known schema
556 language for XML documents. All the markup items it defines are fully compatible with the HTML5
557 specifications (Hickson et al., 2014).

⁶⁵<http://schema.org>

⁶⁶<https://rawgit.com/essepuntato/rash/master/documentation/index.html#metadata>

⁶⁷<https://rawgit.com/essepuntato/rash/master/papers/rash-peerj2016.html>

⁶⁸The full project is available at <https://github.com/essepuntato/rash/>. Please use the hashtag #rashfwk for referring to any of the items defined in the RASH Framework via Twitter or other social platforms.

⁶⁹<http://opensource.org/licenses/ISC>

⁷⁰<http://creativecommons.org/licenses/by/4.0/>

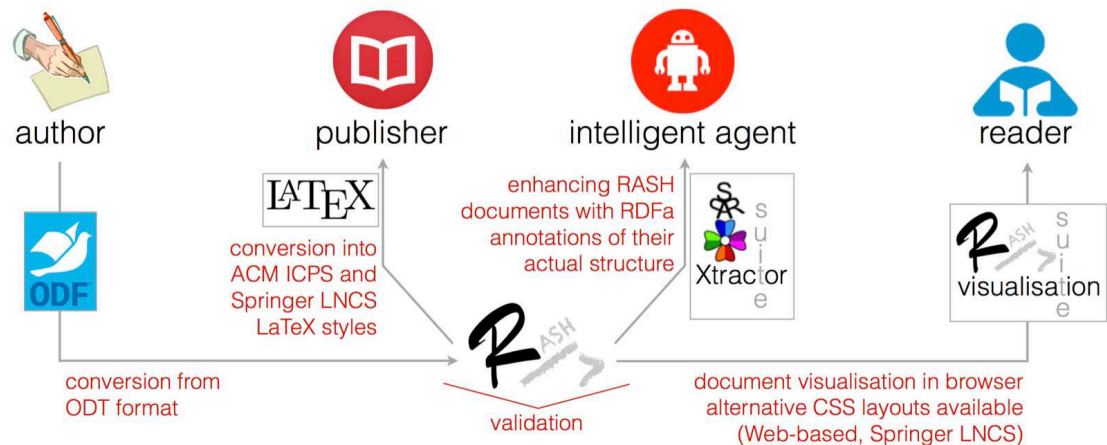


Figure 1. The RASH Framework and its main components.

558 In order to check whether a document is compliant with RASH, we have developed a script to enable
 559 RASH users to check their documents simultaneously both against the specific requirements in the RASH
 560 RelaxNG grammar and also against the full set of HTML checks that the W3C Nu HTML Checker⁷¹
 561 (a.k.a., HTML5 validator) does for all HTML documents (by checking all requirements given in the
 562 HTML specification). This ensures that RASH users get alerted to more potential mistakes in their
 563 documents so that they can easily fix them. Among other things above just using the RASH grammar
 564 only, this script adds relatively sophisticated checking of the datatype microsyntaxes of attribute values.

565 Visualising RASH documents

566 The whole visualisation of this document (as any other RASH document) is rendered by the browser in
 567 the current form by means of appropriate CSS⁷² stylesheets (Atkins J et al., 2015) and Javascript scripts
 568 developed for this purpose.

569 We are actually using some external libraries, i.e., Bootstrap⁷³ and JQuery⁷⁴, in order to guarantee
 570 the current clear visualisation and for adding additional tools to the user. For instance, the footer with
 571 statistics about the paper (i.e., number of words, figures, tables and formulas) and a menu to change the
 572 actual layout of the page⁷⁵, the automatic reordering of footnotes and references, the visualisation of the
 573 metadata of the paper, etc.

574 Note that these kinds of automatic rendering of paper items, such as references to a bibliographic
 575 entry or a figure, reduce the cognitive effort of an author when writing a RASH paper. For instance, a
 576 piece of text referencing a table, e.g. “as shown in Table 2” is created without caring about the particular
 577 text to specify for that reference (“Table 2” in the example), since RASH prescribes to specify just an
 578 empty link to the object one wants to refer to, as shown in the following excerpt:

579 `<p>... as shown in ...</p>`

580 For these objects, the Javascript scripts developed will take care about deciding what is the more
 581 suitable text to put there according to the type of the item referenced.

582 Converting RASH into LaTeX styles

583 We have spent some effort in preparing XSLT 2.0 documents (Kay, 2007) for converting RASH docu-
 584 ments into different LaTeX styles, such as ACM ICPS⁷⁷ and Springer LNCS⁷⁸, among the others. This
 585 is, actually, one of the crucial step to guarantee the use of RASH within international events and to be
 586 able to publish RASH documents in the official LaTeX format as required by the organisation committee

⁷¹<http://validator.w3.org/nu/>

⁷²<http://www.w3.org/Style/CSS/specs.en.html>

⁷³<http://getbootstrap.com/>

⁷⁴<http://jquery.com/>

⁷⁵The layouts currently available are Web-based and Springer’s Lecture Note in Computer Science⁷⁶.

⁷⁷<http://www.acm.org/sigs/publications/proceedings-templates>

⁷⁸<http://www.springer.com/computer/lncs?SGWID=0-164-6-793341-0>

587 of such events. Obviously the full adoption of RASH, of any other Web-first format, would make these
 588 stylesheets not necessary but, currently, they are fundamental for the adoption of the overall approach.

589 Producing RASH from ODT

590 In addition, we have already developed another XSLT 2.0 document to perform conversions from OpenOf-
 591 fice documents into RASH documents, which allows us to write a paper through the OpenOffice editor
 592 and then converting the related ODT file into RASH automatically.

593 The RASH documentation provides a detailed description of how to use OpenOffice for writing sci-
 594 entific documents that can be easily converted in the RASH format. The standard OpenOffice features
 595 (e.g. styles, document properties, etc.), elements (e.g. lists, pictures, captions, footnotes, hyperlinks,
 596 etc.) and facilities (e.g. mathematical editor, cross-reference editor, etc.) can be used to produce fully
 597 compliant RASH documents. A web-based service (for converting documents online) and a Java appli-
 598 cation (that can be downloaded and used offline on the local machine) have been developed to facilitate
 599 the conversion process of OpenOffice documents to the RASH format.

600 In the past few years, as sort of alpha-testing, we have used these conversion approaches with many
 601 internal projects in the Digital and Semantic Publishing Laboratory of the Department of Computer Sci-
 602 ence and Engineering at the University of Bologna. Moreover, also our co-authors and collaborators
 603 from different disciplines (e.g. business and management, humanities, medicine, etc.) have successfully
 604 used this approach for producing their documents, giving us a chance to have fruitful feedbacks, com-
 605 ments and suggestions. In particular, we have been able to convert several ODT files of the main part of
 606 the research papers, project proposals and deliverables, documentation, and two Ph.D. thesis we wrote
 607 in our group into RASH, with a discrete success.

608 ROCS

609 We created an online conversion tool called *ROCS (RASH Online Conversion Service)*⁷⁹ (Di Iorio et
 610 al., 2016) for supporting authors in writing RASH documents and preparing submissions to be easily
 611 processed by current journals, workshops and conferences. ROCS integrates the tools introduced in the
 612 previous sections.

613 The abstract architecture of the tool is shown in Fig. 2. ROCS allows converting an ODT docu-
 614 ment, written according to specific guidelines, into RASH and, then, into LaTeX according to either the
 615 Springer LNCS or the ACM IPCS layouts. Such ODT guidelines⁸⁰ are very simple and use only the
 basic features available in OpenOffice Writer, without using any external tool or plug-in.

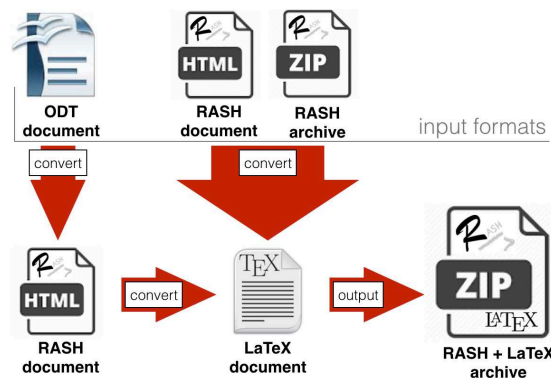


Figure 2. The architecture of ROCS.

616 ROCS allows users to upload three kinds of file, i.e., an ODT document, a HTML file compliant
 617 with RASH, and a ZIP archive which contains an HTML file compliant with RASH and related files (i.e.,
 618 CSSs, javascript files, fonts, images). It returns a ZIP archive containing the original document plus all
 619 its converted versions, i.e., RASH, if an ODT file was given, and the LaTeX file.

620 The main advantage of having the paper both in RASH and in LaTeX is that it is very easy for RASH
 621 to be adopted by workshops, conferences or journals. Since, the program committee, the reviews and
 622

⁷⁹<http://dasplab.cs.unibo.it/rocs>

⁸⁰<https://rawgit.com/essepuntato/rash/master/documentation/rash-in-odt.odt>

623 the editors will also have access to a LaTeX or a PDF version of the paper, the RASH file is an addition
 624 that does not preclude any current workflows. Of course, the hope is that the inherent advantages of an
 625 HTML-based format such as RASH will eventually persuade stakeholders to adopt the HTML version
 626 whenever it will be possible, keeping the alternatives as fallback options.

627 **Enriching RASH documents with structural semantics**

628 A recent development of the RASH Framework has concerned the automatic enrichment of RASH doc-
 629 uments with RDFa annotations defining the actual structure of such documents in terms of the FRBR-
 630 aligned Bibliographic Ontology (FaBIO)⁸¹ and the Document Component Ontology (DoCO)⁸² (Con-
 631 stantin et al., 2016). More in detail, we developed a Java application called *SPAR Xtractor suite*⁸³. SPAR
 632 Xtractor is designed as a one-click tool able to add structural semantics to a RASH document automat-
 633 ically. In fact, SPAR Xtractor takes a RASH document as input and returns a new RASH document
 634 where all its markup elements have been annotated with their actual structural semantics by means of
 635 RDFa. Namely, the tool associates a set of FaBIO or DoCO types with specific HTML elements. The set
 636 of HTML elements and their associations with FaBIO or DoCO types can be customised according to
 637 specific needs of expressivity. The default association provided by the current release of SPAR Xtractor
 638 is the following:

- 639 • the root `html` element is mapped to an individual of the class `fabio:Expression`⁸⁴. The
 640 class `fabio:Expression` identifies the specific intellectual or artistic form that a work takes
 641 each time it is realised;
- 642 • the `body` element is mapped to an individual of the class `doco:BodyMatter`⁸⁵. The class
 643 `doco:BodyMatter` represents the specific intellectual or artistic form that a work takes each
 644 time it is realised;
- 645 • `p` elements are represented as individuals of the class `doco:Paragraph`⁸⁶, i.e. self-contained
 646 units of discourse that deal with a particular point or idea;
- 647 • `figure` elements containing the element `img` within a paragraph are represented as individuals
 648 of the class `doco:FigureBox`⁸⁷, which is a space within a document that contains a figure and
 649 its caption;
- 650 • `section` elements are mapped to individuals of the class `doco:Section`⁸⁸, which repre-
 651 sents logical division of the text. As sections can be organised according to a variable level of
 652 nested sub-sections. Accordingly, SPAR Xtractor reflects this structural behaviour by represent-
 653 ing the containment relation by means of the object property `po:contains`⁸⁹⁹⁰. For example,
 654 a certain `section` element with a nested `section` element produces two individuals of the
 655 class `doco:Section` (e.g. `:section_outer` a `doco:Section` and `:div_inner` a
 656 `section:Section`) related by the property `po:contains` (e.g. `div_outer po:con-`
 657 `tains :div_inner`).

658 In addition to these semantic annotations, which come from the actual structure of a document,
 659 the tool is also able to automatically detect sentences and represent them as individuals of the class
 660 `doco:Sentence`⁹¹. A `doco:Sentence` denotes an expression in natural language forming a single
 661 grammatical unit. For the sentence detection task SPAR Xtractor relies on the sentence detection module
 662 of the Apache OpenNLP project⁹², which provides a machine learning based toolkit for the processing

⁸¹<http://purl.org/spar/fabio>

⁸²<http://purl.org/spar/doco>

⁸³The source code and binaries of SPAR Xtractor are available at <https://github.com/essepuntato/rash/tree/master/sources/spar-xtractor> and <https://github.com/essepuntato/rash/tree/master/tools/spar-xtractor>, respectively.

⁸⁴<http://purl.org/spar/fabio/Expression>

⁸⁵<http://purl.org/spar/doco/BodyMatter>

⁸⁶<http://purl.org/spar/doco/Paragraph>

⁸⁷<http://purl.org/spar/doco/FigureBox>

⁸⁸<http://purl.org/spar/doco/Section>

⁸⁹<http://www.essepuntato.it/2008/12/pattern#contains>

⁹⁰The prefix `po:` stands for the namespace <http://www.essepuntato.it/2008/12/pattern#>.

⁹¹<http://purl.org/spar/doco/Sentence>

⁹²<https://opennlp.apache.org/>

663 of natural language text. By default, SPAR Xtractor is released to support english only. However, it is
 664 possible to extend it with new languages by adding their corresponding models for Apache OpenNLP,
 665 most of which are available with open licence⁹³.

666 We remark that the object property `po:contains` is used for representing any kind of contain-
 667 ment relation among the structural components that SPAR Xtractor deals with. Hence, the usage of
 668 such a property is not limited to the individuals of the class `doco:Section` only. In fact, the prop-
 669 erty `po:contains` can be used, for example, for expressing the containment relation between a
 670 `doco:BodyMatter` and a `doco:Section` or between a `doco:Section` and a `doco:Section`.
 671 For example, let us consider the following code snippets that provides a sample HTML document.

```
672 <html>
673 ...
674 <body>
675 ...
676 <section><h1>A section</h1>
677 ...
678 <p>This is a sentence. This is another sentence of this paragraph.</p>
679 ...
680 <section><h1>A sub-section</h1> ... </section>
681 ...
682 </section>
683 ...
684 </body>
685 </html>
```

686 The HTML document in the snippet above is enriched by SPAR Xtractor resulting in the document
 687 reported in the snippet below.

```
688 <html
689   resource="expression"
690   typeof="http://purl.org/spar/fabio/Expression">
691   ...
692   <body resource="body"
693     typeof="http://purl.org/spar/doco/BodyMatter"
694     property="http://www.essepuntato.it/2008/12/pattern#contains">
695     ...
696     <section resource="section_outer"
697       typeof="http://purl.org/spar/doco/Section"
698       property="http://www.essepuntato.it/2008/12/pattern#contains">
699       <h1 resource="section_outer/title"
700         typeof="http://purl.org/spar/doco/SectionTitle" >
701         <span property="http://purl.org/spar/c4o/hasContent">
702           A section
703         </span>
704       </h1>
705       ...
706       <p resource="section_outer/paragraph-1"
707         typeof="http://purl.org/spar/doco/Paragraph"
708         property="http://www.essepuntato.it/2008/12/pattern#contains" >
709         <span property="http://www.essepuntato.it/2008/12/pattern#contains"
710           resource="section_outer/paragraph-1/sentence-1"
711           typeof="http://purl.org/spar/doco/Sentence">
712         <span property="http://purl.org/spar/c4o/hasContent">
713           This is a sentence.
714         </span>
715       </span>
716       <span property="http://www.essepuntato.it/2008/12/pattern#contains"
717         resource="section_outer/paragraph-1/sentence-2"
718         typeof="http://purl.org/spar/doco/Sentence">
719       <span property="http://purl.org/spar/c4o/hasContent">
720         This is another sentence of this paragraph.
721       </span>>
722     </span>>
723   </p>
724   ...
725   <section resource="section_inner"
726     typeof="http://purl.org/spar/doco/Section"
727     property="http://www.essepuntato.it/2008/12/pattern#contains">
728     <h1 resource="section_inner/title"
729       typeof="http://purl.org/spar/doco/SectionTitle" ">
730       <span property="http://purl.org/spar/c4o/hasContent">
731         A sub-section
732       </span>
733     </h1>
734     ...
```

⁹³Some models are already available under the terms of the Apache Licence at <http://opennlp.sourceforge.net/models-1.5/>.

```
735     </section>
736     ...
737 </section>
738     ...
739 </body>
740 </html>
```

741 RASH AND SAVE-SD: AN EVALUATION

742 The true validation for RASH as a format for research papers rests on its **use by a good number of**
743 **authors and workshops** and its **integration in the publishing process**. For this reason, RASH was
744 first released in conjunction with the Semantics, Analytics, Visualisation: Enhancing Scholarly Data
745 (SAVE-SD 2015) workshop⁹⁴, co-located with WWW 2015. It was subsequently adopted by a number
746 of workshops and conferences⁹⁵. In this section, we will present an evaluation of RASH based on the
747 analysis of questionnaires completed by authors and reviewers of SAVE-SD 2015 and SAVE-SD 2016⁹⁶
748 workshops and a study on RDF annotations in the relevant papers.

749 The users were asked to fill a survey which included a section about their background, a SUS ques-
750 tionnaire and six open questions about their experience with RASH. We will first introduce the two
751 workshops and then discuss and compare the evaluation results. Finally, we will present an analysis
752 of the most frequent vocabularies and entities in RASH papers. The completed questionnaires and the
753 outcomes of the analysis are available at (Osborne and Peroni, 2016).

754 It is worth anticipating that in 2015 there were no converters in the RASH framework, and ROCS
755 was introduced immediately before SAVE-SD 2016. Thus, in both years authors wrote RASH papers
756 with plain text-editors or XML editors, apart from one author that used ROCS in 2016. In general, the
757 authors appreciated RASH and the tools in the RASH framework, even if the editing environment and
758 the converters are still limited.

759 SAVE-SD 2015 and 2016

760 SAVE-SD 2015 was organized by some of the authors of this paper with the aim of bringing together
761 publishers, companies and researchers in order to bridge the gap between the theoretical/academic and
762 practical/industrial aspects in regards to scholarly data. It was thus an inherent multifaceted workshop
763 which drew researchers from a number of heterogeneous fields, such as Document and Knowledge En-
764 gineering, Semantic Web, Natural Language Processing, Scholarly Communication, Bibliometrics and
765 Human-Computer Interaction. Since many of the interested researchers were keen to experiment with
766 novel technologies regarding semantic publishing it was a natural choice for the debut of RASH. For this
767 reason, SAVE-SD 2015 allowed authors to submit papers using either RASH or PDF, explicitly encour-
768 aging authors to try the new format. To this end, the organisers introduced a special award for the best
769 submission in RASH, according to the quality of the markups, the number of RDF statements defined in
770 RDFa, and the number of RDF links to LOD datasets. The possibility of submitting in RASH was also
771 advertised on social media (e.g., Twitter⁹⁷, Facebook⁹⁸) and during various international events (e.g., DL
772 2014⁹⁹, EKAW 2014¹⁰⁰, FORCE 2015¹⁰¹).

773 The initiative had a substantial success: the workshop received 6 out of 23 submissions in RASH
774 and after the review process an additional author chose to prepare the camera ready paper in RASH. Out
775 of these 7 final submissions, 3 were research papers, 1 was a position paper, and 3 posters/demo. These
776 papers were submitted by 16 authors from Switzerland, Italy, Germany, Netherlands, United Kingdom,
777 Ireland, and USA.

778 At the time of the workshop submission deadline, there were no public tools available for convert-
779 ing other formats into RASH. However, the authors were able to self-learn it by simply referring to the
780 documentation page, confirming that computer scientists have no particular problem in handling it di-
781 rectly. The conversion of the RASH submissions into the ACM format requested by Sheridan publisher

⁹⁴<http://cs.unibo.it/save-sd/2015/index.html>

⁹⁵<https://github.com/essepuntato/rash/#rash-papers-accepted-in-scholarly-venues>

⁹⁶<http://cs.unibo.it/save-sd/2016/index.html>

⁹⁷<https://twitter.com/savesdworkshop>

⁹⁸<https://www.facebook.com/savesdworkshop>

⁹⁹<http://www.city.ac.uk/digital-libraries-2014>

¹⁰⁰<http://www.ida.liu.se/conferences/EKAW14/home.html>

¹⁰¹<https://www.force11.org/meetings/force2015>

782 (responsible for the publications of all WWW proceedings) was handled by the organisers through a
 783 semi-automatic process. In particular, they used the XLST files introduced in Section and had to fix
 784 only a few layout misalignments.

785 Six authors and four reviewers involved in SAVE-SD 2015 participated to our evaluation.

786 SAVE-SD 2016 was the second edition of the workshop and had the same characteristics and goals
 787 of the predecessor. In order to give authors full freedom, the organizer decided to accept not only RASH,
 788 but any kind of HTML-based format. Since it was not possible to handle the conversion of any possible
 789 HTML-based format to the publisher layout, the authors of alternative formats were asked to prepare a
 790 PDF of the camera ready version according to the publisher needs.

791 SAVE-SD 2016 received 6 out of 16 submissions in RASH from 14 authors from Italy, Sweden,
 792 Greece, Germany, Belgium, and United States. In total, 5 out of the 14 accepted papers were in RASH,
 793 including 2 full papers, 2 demos and 1 position papers. Even if no author chose to submit in other
 794 HTML-based formats, this possibility will be kept open in future editions. Differently from the previous
 795 edition, the proceedings were published as a dedicated LNCS volume. The conversions of RASH papers
 796 to the PDF documents in Springer LNCS layout was automatically handled by ROCS.

797 As in the previous edition, we evaluated RASH by conducting the same study (with the same exact
 798 questions). Seven authors of RASH papers and three reviewers participated to the survey.

799 User background

800 It is useful to first assess the background of RASH pioneer users in term of their knowledge of relevant
 801 technologies and software. For this reason, the first section of the survey included a number of statements
 802 about the user expertise (e.g., "I have extensive experience in writing academic papers with LaTeX ")
 803 and allowed five response options, from "Strongly Agree" to "Strongly Disagree". Table 3 shows the
 804 percentage of users who claimed to be familiar with a range of technologies (by selecting "Agree" or
 805 "Strongly Agree").

Table 3. User background for SAVE-SD 2015, SAVE-SD 2016, and average values.

Year	MS Word	OO Writer	LaTeX	HTML	XML	RelaxNG	SW	RDFa	Turtle	JSON-LD
2015	33%	33%	83%	83%	100%	67%	83%	100%	100%	50%
2016	57%	0%	71%	71%	71%	29%	57%	57%	57%	43%
AVR	40%	13%	67%	67%	73%	40%	60%	67%	77%	40%

806 In 2015, the authors were mainly from the Semantic Web community and therefore familiar with
 807 technologies such as RDFa and Turtle. Most of them knew how to correctly annotate a HTML file
 808 and understood the advantages of including semantic relationships in the paper. They also commonly
 809 used LaTeX rather than Microsoft Word or OpenOffice Writer. This suggests that they were acquainted
 810 with WYSIWYG editors and had experience with complex formats. A qualitative analysis of the survey
 811 answers confirms this intuition; for example an author remarked: "I am used to writing papers in LaTeX
 812 so I do not want to bother with formatting and in that sense RASH is similar".

813 In 2016 the situation changed and only 57% of the users were familiar with semantic technologies.
 814 In addition, even if most of them knew how to use LaTeX, the majority of them had experience also with
 815 Microsoft Word. It seems thus that RASH started to interest also less technical users with a different
 816 research backgrounds.

817 User survey

818 We assessed strengths and weaknesses of RASH by means of six open questions. We summarize here
 819 the answers of both authors and reviewers for the 2015 and 2016 edition. The reviewers answered only
 820 questions 2, 3, 4 and 5. Note that the questions were exactly the same in both editions and none of the
 821 participants filled both the surveys.

822 SAVE-SD 2015 Survey

- 823 • [Q1] Why did you choose the RASH format for your paper?

824 Four authors answered that the main reason was to try it out, mostly because they "supported the
 825 idea of publishing academic papers as HTML" and were convinced that "PDF should be replaced".

- 826 Two of them added that they were motivated by the possibility of adding semantic annotations to
827 their papers.
- 828 • [Q2] *How effectively did RASH support you in writing/reviewing the paper?*
829 The majority of the authors suggested that some tasks, such as setting up the bibliography, were
830 still cumbersome. They added that the development of tools that could solve these issues and hide
831 the technical details to the common users would be very important for a broader adoption. The
832 reviewers remarked that their experience was very similar to reviewing a paper in PDF format and
833 did not present any particular challenge (e.g., “did not have many features that would distinguish
834 it from a PDF”, “it met all of my needs and was easy to use”).
 - 835 • [Q3] *What were the most useful features of RASH to help you writing/reviewing the paper?*
836 The authors listed a number of functionalities including the multiple graphical layouts (2 authors),
837 the support of RDFa annotations (2) and the built-in validation (1). The ability to display the paper
838 according to different layouts was also praised by reviewers.
 - 839 • [Q4] *What were the main weaknesses that RASH exhibited in supporting the writing/reviewing of
840 the paper?*
841 Most authors suggested that the handling of bibliography, figures and captions should be improved.
842 Half of them also pointed out that the manual insertion of semantic annotations was cumbersome
843 and a large amount of RDFa “introduces a bit of confusion in the paper”. An author observed that
844 using the word count as limit in the RASH venues rather than the number of pages introduces the
845 issue of possibly exceeding the editor limits. Most reviewers did not report any problem in using
846 RASH for assessing a paper. However, one of them noted that it still lacked a menu for easily
847 navigating the different sections, as PDF files instead support.
 - 848 • [Q5] *Can you think of any additional features to be included in RASH that would have helped you
849 to write/review the paper?*
850 The majority of authors suggested that the aforementioned limitations were mainly due to the use
851 of a HTML editor and it will be imperative to develop a WYSIWYG editor or a tool to convert
852 from ODT to RASH. A user also suggested developing a tool for graphically showing the semantic
853 annotations, as “what is linked to what, in order to check the correctness of assertions” and a
854 reviewer advised to implement a way to easily access to the different sections of the document.
 - 855 • [Q6] *Would you use RASH regularly for writing your academic papers?*
856 Five out of six authors answered they would like to keep using RASH. Most of them however
857 added that this would also depend on the creation of a better editor and a solid array of tools for
858 managing technical details and converting standard formats for writing research paper to and from
859 RASH.

860 **SAVE-SD 2016 Survey**

- 861 • [Q1] *Why did you choose the RASH format for your paper?*
862 As with the 2015 results, the majority of the authors (4) claimed that they adopted it for trying a
863 new format, three authors because they were motivated by the workshop and three because they
864 actively support the ideas behind RASH.
- 865 • [Q2] *How effectively did RASH support you in writing/reviewing the paper?*
866 Five users wrote the papers directly in RASH and only one used Open Office and then converted
867 it with ROCS. In the first group, one user was positive, one neutral, and three suggested the need
868 for a WYSIWYG editor, since “writing in html is not so effective” and “not everyone [of the co-
869 authors] knew how to validate against the schema”. In particular, it was suggested the need for a
870 Microsoft Word converter, since the ODT produced by Microsoft Word could not be processed by
871 ROCS. As in 2015, the reviewers did not find many differences with respect to PDF papers. One
872 of them claimed to actually prefer RASH since it “makes better use of the page space”.
- 873 • [Q3] *What were the most useful features of RASH to help you writing/reviewing the paper?*
874 The authors mentioned a variety of different features including the formatting semantics (“no
875 worries about section and layout”), the bibliographic reference management and the ability to
876 display the paper according to different layouts. A reviewer also praised the ability of converting
877 RASH to PDF.
- 878 • [Q4] *What were the main weaknesses that RASH exhibited in supporting the writing/reviewing of
879 the paper?*
880 Differently with 2015, the authors had no particular problem with the handling of bibliography,

881 figures and captions. However, most of authors (5) pointed to the fact that with the current version
 882 they still had to write HTML, which is usually not straightforward. Three of them suggested
 883 solving the problem by introducing a WYSIWYG editor, while two of them suggested creating
 884 new converters to translate LaTeX and Microsoft Word into RASH. One user also flagged that
 885 the visualization of RASH document can change in different browsers. The reviewers, as in 2015, did
 886 not report any particular problem in using RASH.

887 • [Q5] Can you think of any additional features to be included in RASH that would have helped you
 888 to write/review the paper?

889 Consistently with the aforementioned weaknesses and the 2015 results the users called for the
 890 creation of a WYSIWYG editor (3) and a way to convert from LaTeX and Microsoft Words (3). In
 891 addition, a user suggested a tool for automatically generating a bibliography, similar to BibTeX.

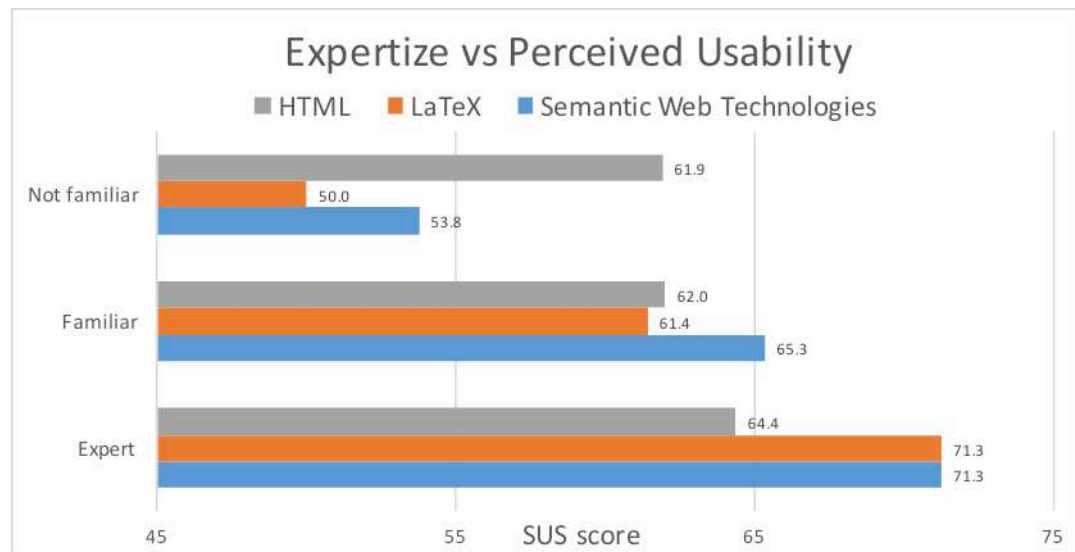
892 • [Q6] Would you use RASH regularly for writing your academic papers?

893 Three authors asserted that they would be happy to keep using RASH even in the current version,
 894 two of them that they were ready to use it again, depending on its development, and only one was
 895 negative about it.

896 **RASH usability**

897 We also performed a quantitative analysis of the usability of RASH, using the System Usability Scale
 898 (SUS) questionnaire (Brooke, 1996). The scores are acceptable, though not very high, especially if we
 899 consider that all authors but one edited RASH files directly with text/XML editors. Users perceived even
 900 a 'vanilla RASH' as acceptable, though they need more sophisticated converters as remarked in the open
 901 questions of the survey.

902 RASH yielded a mean score of 62.7 ± 11.9 , slightly lower than the average SUS score (68). However,
 903 SUS scores varied dramatically according to the person background. Fig. 3 shows the results of different
 904 categories of expertise¹⁰² in HTML, LaTeX and Semantic Web Technologies (SWT), which appear
 905 correlated with the average SUS scores (respectively $r = 0.78, 0.97, 0.99$). Authors with a strong ex-
 906 pertize in LaTeX and SWT yielded significantly better SUS scores, while authors with HTML expertise
 907 yielded only slightly better scores. For this reason, authors from 2015, who as previously discussed had a
 908 higher expertise in these categories obtained an average SUS score of 69.6 ± 11.9 , while the authors from
 909 2016 yielded 57.1 ± 9.7 . However, the difference is not statistically significant because the two samples
 are small and the test power low.



910 **Figure 3.** User expertise in HTML, LaTeX and Semantic Web Technologies versus average SUS score.

911 These results further confirm that most users with limited expertise in non-WYSIWYG editors and
 912 semantic technologies find unfeasible to write HTML directly, even in a simplified form.

¹⁰²The authors who answered "Strongly Agree" to the background questions were classified as "Experts", the ones who answered "Agree" as "Familiar", and all the others as "Not familiar".

913 Analysis of RDF annotations in RASH documents

914 To complete the previous analysis, we also studied the nature of the semantic annotations in RASH
 915 papers. We focused on a sample of 1751 annotations obtained from 11 papers published in SAVE-SD
 916 2015 and 2016. The number of statements in a single paper was found to range from 24 to 903, yielding
 917 a median value of 46 (25th percentile 34, 75th percentile 175). We extracted all the RDF statements by
 918 running the W3C RDFa 1.1 Distiller service¹⁰³ on each article. We then considered only the statements
 919 that used http-based entities as predicates, or their objects if used for typing resources. The data are
 920 organised in several CSV files and have been obtained by running a Python script we developed for
 921 gathering the data used in this evaluation.

922 The first goal of the study was to determine the prevalent vocabularies and how much they were used
 923 in the average paper. The left panel of Fig. 4 shows the common vocabularies. Schema.org and PRISM
 924 are actually enforced by RASH: the first is used for standard metadata such as emails, affiliations and
 925 organization names and the second for keywords. In addition, a quantity of XHTML statements was
 926 automatically extracted when processing DPUB-ARIA roles (Garrish et al., 2016). Thus we will not
 927 consider such vocabularies in the rest of the evaluation. The other common vocabularies are Dublin
 928 Core, which appear in 82% of the papers, FOAF (27%) and the SPAR ontologies (Peroni, 2014), such
 929 as FABIO (36%) and CITO (27%) (Peroni and Shotton, 20122012). The right panel of Fig. 4 illustrates
 930 the average number of statement for each of these vocabularies. Dublin Core characterizes the highest
 number of annotation (9.4), followed by FOAF (7.4) and FABIO (6.4).

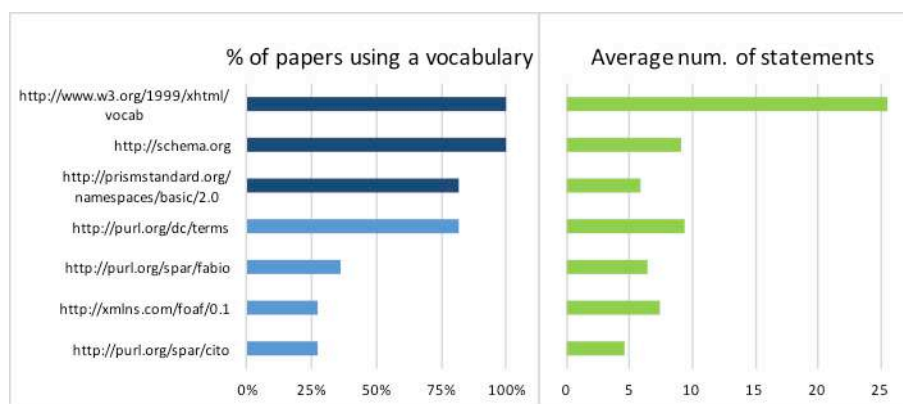


Figure 4. Percentage of papers and average number of statements using a vocabulary

931

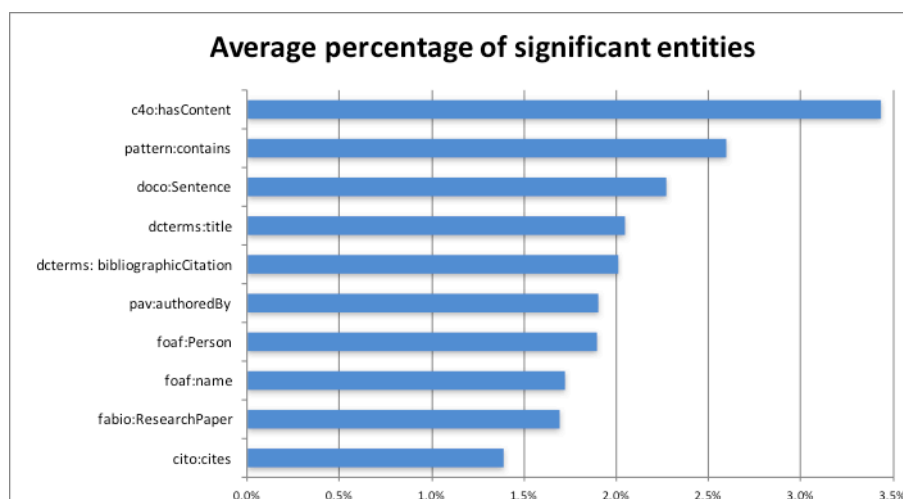


Figure 5. Average percentage of vocabulary entities in a RASH paper (excluding the mandatory ones).

¹⁰³<https://www.w3.org/2012/pyRdfa/>

932 We also performed a more fine-grained analysis considering the amount of entities of these vocabularies within the various RDF statements. The goal was to understand the percentage of contribution that the various entities provide (on average) to the statements of the document analysed. As expected, 933 the entities that contribute to about 60% of the statements are either those that are obliged by RASH 934 (prism:keyword 6.9%, schema:affiliation 5.7%, schema:name 5.3%, and schema:email 935 4.7%) or those automatically extracted by processing the DPUB roles included, mandatorily, in the documents (xhtml:role 38%). Excluding these, the following top ten entities, shown in Fig. 5, cover 936 about 20% of the statements. 937

938 Among these entities, we have three classes describing three diverse but interlinked kinds of objects, i.e. people (foaf:Person) authoring a research work (fabio:ResearchPaper) and the sentences (doco:Sentence) therein contained. The other seven entities are three object properties 939 – two of them (pav:authoredBy and pattern:contains) provide the links between the three 940 aforementioned classes, while the other, i.e., cito:cites, describes citation links between papers – 941 and four data properties – used for providing additional metadata about the entities (dcterms:title, 942 dcterms:bibliographicCitation, foaf:name) and for describing bunches of textual content 943 of the sentences (c4o:hasContent). 944

945 Discussion

946 The evaluation study confirmed that RASH can already be adopted in workshops, conferences and journals and can be quickly learnt by researchers who are familiar with HTML. However, it also highlighted 947 some issues in the adoption of HTML formats, especially by less technical savvy users.

948 The 2016 survey showed that RASH is currently being tried also by users unfamiliar with semantic web technology. While the expansion of the user base represents a positive development, it also yields a 949 number of challenges. The mass of authors accustomed to WYSIWYG editors such as Microsoft Word 950 or OpenOffice Writer, tend to have difficulties with HTML editors. In addition, since research papers 951 are often written by multiple authors, it is usually simpler to use the most well-known solutions. For 952 these reasons, we need to offer the authors who cannot or do not want to change their workflow the tools 953 for converting their favourite format to RASH and annotate the resulting paper. While ODT was a first 954 step in this direction, it is imperative to be also able to process DOCX and LaTeX. A second important 955 issue is that authors who are not expert in semantic technologies can find hard to correctly annotate their 956 papers. Hence, we also need to develop simple tools for helping authors in this phase. The introduction 957 of these solutions will be critical for motivating users to adopt HTML-based approaches and for creating 958 a robust framework that could be used by expert and common users alike.

959 As far it concerns the analysis of RDF annotations in RASH documents, the outcomes highlighted 960 that the users spontaneously decided to adopt few well-known standard vocabularies, rather than using 961 a multiplicity of different solutions. The most used vocabularies other than Schema.org and PRISM, 962 which are enforced by default in RASH, are Dublin Core, FOAF, and the SPAR ontologies. However, 963 the outcomes of our evaluation generally show a quite low number of statements specified by the authors. 964 This behaviour could derive from the lack of appropriate support for the annotation of RASH papers with 965 RDF data. In addition, this low number seems not to be related to the research community the authors 966 work in. For instance, several of the papers written by Semantic Web experts do not include any RDF 967 statement in addition to those annotations that are enforced by RASH. 968

969 CONCLUSIONS

970 In this paper we have introduced *RASH*, a markup language defined as a subset of HTML for writing 971 scientific articles, and the *RASH Framework*, a set of specifications and tools for writing articles in RASH. 972 In particular, we have discussed the rationale behind the development of RASH, and we have presented 973 the language and the validation/visualisation/conversion/extraction tools developed so far.

974 The goal of the paper was also to investigate the applicability and the potentialities of RASH, though 975 the evaluation of its adoption in two editions of the SAVE-SD workshop. To the best of our knowledge, 976 this is the first empirical evaluation on the adoption of HTML-based languages for writing scientific 977 papers. The experiments proved that RASH can be successfully used for workshops and conferences, 978 with a good acceptance by the authors and a smooth integration in the existing publishing process. 979

980 As immediate future developments, we plan to develop tools for automating the process of semantic 981 enrichment of RASH documents. For instance, we are currently working on the automatic identification 982

985 of section rhetorics and citation functions so as to describe them according to two SPAR Ontologies
 986 (Peroni, 2014), i.e. the Document Component Ontology (DoCO)¹⁰⁴ and the Citation Typing Ontology
 987 (CiTO)¹⁰⁵ respectively. We also are developing a Web tool for validating RASH document so as to spot
 988 possible syntactic errors easily.

989 We also intend to further develop the RASH framework. Firstly, we are working on more sophisti-
 990 cated authoring tools and converters. For instance, we are currently developing additional XSLT docu-
 991 ments in order to convert DOCX documents into RASH and to convert RASH documents into several
 992 different LaTeX formats for scholarly communications – such as IEEE conference proceedings, ACM
 993 journals, IOS Press journals and PeerJ – as well as into EPUB for easing its (offline) portability in mobile
 994 devices.

995 We are also studying the possibility of integrating AsciiMath¹⁰⁶ as additional language for writing
 996 formulas and correctly translating it in other standard formats (e.g. MathML and LaTeX) so as to handle
 997 it during the conversion process available in ROCS. In addition, we are experimenting techniques for
 998 automatically generating accessible graphs from data contained in a referenced CSV file.

999 **Acknowledgements.**

1000 We would like to thank Sarven Capadislil¹⁰⁷ for our inspiring discussions on the topic, all the authors and
 1001 the reviewers of the accepted papers of the SAVE-SD 2015¹⁰⁸ and the SAVE-SD 2016¹⁰⁹ workshops for
 1002 having provided us useful suggestions and insights for improving RASH and the related tools, as well as
 1003 all the other early adopters of RASH. We would also like to thank the other two organisers of the past
 1004 two edition of SAVE-SD, i.e. Jun Zhao¹¹⁰ and Alejandra Gonzalez-Beltran¹¹¹ for supporting us in the
 1005 adoption of RASH as possible HTML submission format. In addition, we are particularly grateful to all
 1006 the GitHub users that suggested and introduced new features to RASH and developed the tools included
 1007 in its Framework: Mike Smith¹¹² and Ruben Verborgh¹¹³.

1008 **REFERENCES**

- 1009 Atkins Jr. T., Etemad E. J., Rivoal F. (2015). CSS Snapshot 2015. W3C Working Group Note 13 October
 1010 2015. World Wide Web Consortium. <https://www.w3.org/TR/css3-roadmap/>
 1011 Berjon R., Ballesteros S. (2015). What is Scholarly HTML? [http://scholarly.vernacular.
 1012 io/](http://scholarly.vernacular.io/)
 1013 Bourne P. E., Clark T., Dale R., de Waard A., Herman I., Hovy E. H., Shotton D. (2011). FORCE11
 1014 White Paper: Improving The Future of Research Communications and e-Scholarship. White paper, 28
 1015 October 2011. FORCE11. https://www.force11.org/white_paper
 1016 Brooke J. (1996). SUS-A quick and dirty usability scale. Usability evaluation in industry 189, no. 194
 1017 (1996): 4-7.
 1018 Capadislil S., Riedl R., Auer S. (2015). Enabling Accessible Knowledge. In Proceedings of the 2015 Inter-
 1019 national Conference for E-Democracy and Open Government (CeDEM 2015). Krems, Austria: Uni-
 1020 versitt Krems. Open Access at <http://csarven.ca/enabling-accessible-knowledge>
 1021 Carlisle D., Ion P., Miner R. (2014). Mathematical Markup Language (MathML) Version 3.0 2nd Edition.
 1022 W3C Recommendation 10 April 2014. World Wide Web Consortium [http://www.w3.org/TR/
 1023 MathML3/](http://www.w3.org/TR/MathML3/)
 1024 Clark J., Makoto M. (2001). RELAX NG Specification. Committee Specification, 3 December 2001.
 1025 OASIS. <http://relaxng.org/spec-20011203.html>
 1026 Constantin A., Peroni S., Pettifer S., Shotton D., Vitali F. (2016). The Document Component Ontol-
 1027 ogy (DoCO). In *Semantic Web*, 7 (2): 167-181. <http://dx.doi.org/10.3233/SW-150177>

¹⁰⁴<http://purl.org/spar/doco>

¹⁰⁵<http://purl.org/spar/cito>

¹⁰⁶<http://asciimath.org/>

¹⁰⁷<http://csarven.ca/>

¹⁰⁸<http://cs.unibo.it/save-sd/2015/accepted-papers.html>

¹⁰⁹<http://cs.unibo.it/save-sd/2016/accepted-papers.html>

¹¹⁰<http://www.oerc.ox.ac.uk/people/jun-zhao>

¹¹¹<http://www.oerc.ox.ac.uk/people/alejandra>

¹¹²<https://sideshowbarker.net/>

¹¹³<http://ruben.verborgh.org>

- 1028 (Open Access at <http://www.semantic-web-journal.net/system/files/swj1016.pdf>)
1029 pdf)
- 1030 Cyganiak R., Wood D., Lanthaler M. (2014). RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation, 25 February 2014. World Wide Web Consortium. <http://www.w3.org/TR/rdf11-concepts/>
1031
1032
- 1033 Di Iorio A., González Beltrán A., Osborne F., Peroni S., Poggi F., Vitali F. (2016). It ROCS!: The RASH Online Conversion Service. In WWW (Companion Volume) 2016: 25-26. <http://dx.doi.org/10.1145/2872518.2889408> (Open Access at <https://rawgit.com/essepuntato/rash/master/papers/rash-poster-www2016.html>)
1034
1035
1036
- 1037 Di Iorio A., Peroni S., Poggi F., Vitali F. (2014). Dealing with structural patterns of XML documents. *Journal of the American Society for Information Science and Technology*, 65(9): 1884–1900. <http://dx.doi.org/10.1002/asi.23088>
1038
1039
- 1040 Diggs J., Craig J., McCarron S., Cooper M. (2015). Accessible Rich Internet Applications (WAI-ARIA) 1.1. W3C Working Draft 19 November 2015. World Wide Web Consortium. <http://www.w3.org/TR/wai-aria-1.1/>
1041
1042
- 1043 Gandon F., Schreiber G. (2014). RDF 1.1 XML Syntax. W3C Recommendation, 25 February 2014. World Wide Web Consortium. <https://www.w3.org/TR/rdf-syntax-grammar/>
1044
- 1045 Gao S., Sperberg-McQueen C. M., Thompson H. S. (2012). W3C XML Schema Definition Language (XSD) 1.1 Part 1: Structures. W3C Recommendation, 5 April 2012. World Wide Web Consortium. <https://www.w3.org/TR/xmlschema11-1/>
1046
1047
- 1048 Garrish M., Siegman T., Gylling M., McCarron S. (2016). Digital Publishing WAI-ARIA Module 1.0. W3C Working Draft, 17 March 2016. World Wide Web Consortium. <https://www.w3.org/TR/dpub-aria-1.0/>
1049
1050
- 1051 Hickson I., Berjon R., Faulkner S., Leithead T., Doyle Navara E., O'Connor E., Pfeiffer S. (2014). HTML5: A vocabulary and associated APIs for HTML and XHTML. W3C Recommendation, 28 October 2014. World Wide Web Consortium. <http://www.w3.org/TR/html5/>
1052
1053
- 1054 JTC1/SC34 WG 4. (2011). ISO/IEC 29500-1:2011 - Information technology - Document description and processing languages - Office Open XML File Formats - Part 1: Fundamentals and Markup Language Reference. Geneva, Switzerland: International Organization for Standardization. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=59575
1055
1056
1057
1058
- 1059 JTC1/SC34 WG 6. (2006). ISO/IEC 26300:2006 - Information technology - Open Document Format for Office Applications (OpenDocument) v1.0. Geneva, Switzerland: International Organization for Standardization. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=43485
1060
1061
1062
- 1063 Kay M. (2007). XSL Transformations (XSLT) Version 2.0. W3C Recommendation, 23 January 2007. World Wide Web Consortium. <http://www.w3.org/TR/xslt20/>
1064
- 1065 Lin T. T. Y., Beales G. (2015). ScholarlyMarkdown Syntax Guide. Guide, 31 January 2015. <http://scholarlymarkdown.com/Scholarly-Markdown-Guide.html>
1066
- 1067 National Information Standards Organization. (2012). JATS: Journal Article Tag Suite. American National Standard No. ANSI/NISO Z39.96-2012, 9 August 2012. http://www.niso.org/apps/group_public/download.php/10591/z39.96-2012.pdf
1068
1069
- 1070 Osborne F., Peroni S. (2016). Outcomes of SAVE-SD 2015 and 2016 questionnaires on RASH and analysis of RDF annotations in RASH papers. Figshare. <http://dx.doi.org/10.6084/m9.figshare.3980463>
1071
1072
- 1073 Peroni S. (2014). *Semantic Web Technologies and Legal Scholarly Publishing*. Law, Governance and Technology Series 15. Cham, Switzerland: Springer. ISBN: 978-3-319-04776-8. <http://dx.doi.org/10.1007/978-3-319-04777-5>
1074
1075
- 1076 Peroni S. (2014). The Semantic Publishing and Referencing Ontologies. In *Semantic Web Technologies and Legal Scholarly Publishing*: 121-193. Cham, Switzerland: Springer. http://dx.doi.org/10.1007/978-3-319-04777-5_5 (Open Access at <http://speroni.web.cs.unibo.it/publications/peroni-2014-semantic-publishing-referencing.pdf>)
1077
1078
1079
- 1080 Peroni S., Lapeyre D. A., Shotton D. (2012). From Markup to Linked Data: Mapping NISO JATS v1.0 to RDF using the SPAR (Semantic Publishing and Referencing) Ontologies. In *Proceeding of the Journal Article Tag Suite Conference 2012 (JATS-Con 2012)*. Bethesda, Maryland, USA: National Center for
1081
1082

- 1083 Biotechnology Information. <http://www.ncbi.nlm.nih.gov/books/NBK100491/>
- 1084 Peroni S., Shotton D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources
1085 and citations. In *Web Semantics*, 17 (December 2012): 33-43. <http://dx.doi.org/10.1016/j.websem.2012.08.001> (Open Access at <http://speroni.web.cs.unibo.it/publications/peroni-2012-fabio-cito-ontologies.pdf>)
- 1086
1087
- 1088 Pettifer S., McDermott P., Marsh J., Thorne D., Villeger A., Attwood T. K. (2011). Ceci n'est pas un
1089 hamburger: modelling and representing the scholarly article. *Learned Publishing*, 24(3): 207-220.
1090 <http://dx.doi.org/10.1087/20110309>
- 1091 Prud'hommeaux E., Carothers G. (2014). Turtle - Terse RDF Triple Language. W3C Recommendation,
1092 25 February 2014. World Wide Web Consortium. <http://www.w3.org/TR/turtle/>
- 1093 Raggett D., Le Hors A., Jacobs I. (1999). HTML 4.01 Specification. W3C Recommendation, 24 Decem-
1094 ber 1999. World Wide Web Consortium. <http://www.w3.org/TR/html401/>
- 1095 Shotton D., Portwin K., Klyne G., Miles A. (2009). Adventures in Semantic Publishing: Exemplar
1096 Semantic Enhancements of a Research Article. *PLoS Computational Biology*, 5(4): e1000361. <http://dx.doi.org/10.1371/journal.pcbi.1000361>
- 1097
1098 Sporny M., Kellogg G., Lanthaler M. (2014). JSON-LD 1.0 - A JSON-based Serialization for Linked
1099 Data. W3C Recommendation, 16 January 2014. World Wide Web Consortium. <https://www.w3.org/TR/json-ld/>
- 1100
1101 Sporny M. (2013). HTML+RDFa 1.1: Support for RDFa in HTML4 and HTML5. W3C Recommen-
1102 dation, 22 August 2013. World Wide Web Consortium. <http://www.w3.org/TR/rdfa-in-html/>
- 1103
1104 Walsh N. (2009). The DocBook Schema Version 5.0. OASIS Standard, 1 November 2009. Burlington,
1105 Massachusetts, US: Organization for the Advancement of Structured Information Standards. <http://docs.oasis-open.org/docbook/specs/docbook-5.0-spec-os.html>
- 1106