Markup Languages Support for Content Management of Agricultural Portals

Jan Masner[1], Jiří Vaněk[2], Jan Jarolímek[2], Vladimír Očenášek[2]

[1]Department of Information Technologies, Czech University of Life Sciences, Czech Republic, e-mail: masner@pef.czu.cz
[2]Department of Information Technologies, Czech University of Life Sciences, Czech Republic

Abstract. Within a wider research, the Department of Information Technologies of CULS Prague works on a project, which deals with methodology for creation, updates, store and presentation of information content in the WWW environment. This paper provides analysis of markup languages and discusses other proposals for the future research and development. The results of this study indicate that the HTML5 language should be used as a main carrier. An abstract envelope (using XML, JSON or both) should be defined to keep information content in blocks to allow portability, communication with CMS, mobile devices or other applications.

Keywords: HTML, WWW, CMS, Information content, WYSYWIG.

1 Introduction

Internet and the World Wide Web environment have rapidly spread over the population during past years. Number of web pages and portals is still growing. Recently, there are more than 1.2 billion of web pages online (Internet Live Stats, 2014). The past decade has seen the rapid departure from classical printed media (newspapers, magazines). At the same time, the number of online media and portals grows. Establishing and managing of such websites is becoming easier. These trends are expected for the future to be more significant. The online publication and online content brings many opportunities (Das et. Al, 2009).

Due to a development of internet technologies, especially content management systems, even users without knowledge of desired technologies (HTML, CSS) can manage the online content (Brown, 2014). This usually means creation of posts like articles, news, interviews, etc. The main part of the content creation utilizes WYSIWYG (What You See Is What You Get) editors. Other tools of content management systems and content parts are connected in addition (photographs, links to related content, external links and attachments in general). WYSIWYG editors allow users to work with the content without knowledge of the desired technologies the same way like any text processors do. However, this suffers from many limitations.

Due to a lack of knowledge of HTML and connected CSS, the work with advanced features (such as floating property) is difficult. WYSIWYG editors and

Proceedings of the 7th International Conference on Information and Communication Technologies in Agriculture, Food and Environment (HAICTA 2015), Kavala, Greece, 17-20 September, 2015.

tools are not perfect and can lead to problems with inconsistency of output. These issues need to be solved programmatically then (Spiesser and Kitchen, 2004). Digital form of online media and content offers an undiscovered potential of opportunities. The textual content can be extended by various components such as images with description, galleries, links, maps and interactive objects in general etc.

There is a need for analysis of the content storage form. Recent content management systems uses various types of WYSIWYG editors, which can be extended by additional elements. There are many modern technologies for content sharing, searching and classification such as semantic web, metadata description (e.g. AGROVOC), sharing (e.g. OAI-PMH), Internet of Things and cooperation with mobile devices (Šimek et. al, 2013; Šimek, Stočes and Vaněk, 2014).

Within a wider research, the Department of Information Technologies of CULS Prague works on a project that deals with methodology for creation, updates, store and presentation of information content in the WWW environment. The methodology should cover modern trends in the informatics area such as usability, User eXperience (UX), Internet of Things (IoT), connection with applications for mobile devices, etc. Finally, it should be generalizable and usable for various types of contents in different areas. Regarding the agrarian sector where IT technologies penetration and knowledge of desired technologies is at a lower level, the importance of the research topic is significant.

The objective of this paper are to determine whether the HTML5 language is optimal to be used in the target methodology. In addition, the study analyses other initial aspects for the research project.

2 Materials and Methods

The following section deals with the analysis of the Hypertext Markup Language (HTML) and its changes during the specifications development. The main objective is to determine, what are the most important changes regarding the information content and what changes are expected for the future. Additionally an overview of content management systems is presented as well.

Tim Berners-Lee formed the HTML in CERN (the European Laboratory for Particle Physics in Geneva) laboratories in 1989. The basic idea was to allow an efficient organisation and availability of scientific documents from remote places. However, instead of making the documents simply available to download, they should have been linked together by hypertext links. Design of the language was based on SGML standard (ISO 8879:1986 Information processing - Text and office systems - Standard Generalized Markup Language). As Ragget (1998) writes: '*The SGML elements used in Tim's HTML included P (paragraph); H1 through H6 (heading level 1 through heading level 6); OL (ordered lists); UL (unordered lists); LI (list items) and various others. What SGML does not include, of course, are hypertext links: the idea of using the anchor element with the HREF attribute was purely Tim's invention, as was the now-famous `www.name.name' format for addressing machines on the Web.*'

The first HTML version was not standardised anywhere and so does not have any version name. The first official version HTML 2 was introduced by IETF (Internet Engineering Task Force) working group in 1994. Later the specification development was moved to World Wide Web Consortium (W3C).

The following chapters analyse HTML specifications based on appropriate specifications by W3C.

2.1 HTML 4.01

The specification version 4.01 was the first widely spread and used version. The W3C recommendation 4.0, later revisioned to 4.01 was released in December 1999 (W3 CONSORCIUM, 1999). In terms of information content, most of the specification is important. Chapter 14 (W3 CONSORCIUM,1999) mentions support for CSS styling. Later, this aspect appeared to be a key factor in defining the look of websites. The specification covers elements for several textual content types. For the research purposes we can define several key chapters and parts of the language:

- 7.5.1 Headings
- 9 Text
- 10 Lists
- 11 Tables
- 12 Links
- 15 Alignment, font styles, and horizontal rules

One of the most important aspects for the study objectives is the language syntax. The HTML till 4.01 version was based on SGML – so it is an application of the standard. Most important rules are:

- Element names are always case-insensitive
- Certain elements were permitted to omit the end tag
- Attribute names are always case-insensitive
- Attribute values can stand without quotes (under certain conditions)
- Some attributes have no value

2.2 XHTML

In 2000 a new specification was introduced. *XHTML 1.0* was released and in 2002 revised. This version is still used on many websites, but is continuously replaced. As W3 CONSORCIUM (2000) writes: '*XHTML is a family of current and future document types and modules that reproduce, subset, and extend HTML 4*' Actually, in terms of information content, the specification is nothing else than reworked HTML 4.01 using rules of XML.

The main differences cover 2 aspects. First, attribute name as identifier for elements *a, applet, form, frame, iframe, img,* and *map* is formally deprecated. The ID reference type from XML should be used instead. Second and more important change is in syntax. Since XHTML is an application of XML instead of SGML, the use of tags, elements and attributes has more limitations. Generally it has to agree with XML syntax and rules. The most important aspects follow:

- Elements and attributes names are always lower case (XML is case-sensitive) - ~~*<BODY>*~~ → *<body>*
- All elements must have end tags - ~~*<br>*~~ → *<br/>*
- Attributes must have a value
- Attribute values must always be quoted - *class=bold* → *class="bold"*
- Overlapping of elements is not permitted (must be well-formed)
  - ~~*<p>here is an emphasized <em>paragraph.</p></em>*~~
  - *<p>here is an emphasized <em>paragraph</em>.</p>*

In 2010, W3C released the specification XHTML Modularization 1.1, which is and abstract modularization of XHTML (W3 CONSORCIUM, 2010). On the basis of this document specification *XHTML™ 1.1 - Module-based XHTML* and *XHTML Basic 1.1 - Second Edition* was created (W3 CONSORCIUM, 2010). The main (and only) contribution was the modular architecture. So it is based on several modules that relies on certain areas. The specification was not widely used probably thanks to already ongoing development of HTML5. The XHTML 1.1 specification is based on Strict variant of 1.0. Regarding the information content, there are only two differences. Firstly, the element *name* was finally removed and is now obsolete. Secondly, element *ruby* has been added.

2.3 HTML5

Originally the development of HTML5 was performed by *WHATWG* group. The first public draft by W3C was released already in 2008 during the finishing of XHTML 1.1 (W3 CONSORCIUM, 2008). Final recommendation was released in 2014 (W3 CONSORCIUM, 2014). Nowadays the HTML5 is the most widely spread version used on websites (Power Mapper, 2015) as shown on Figure 1.

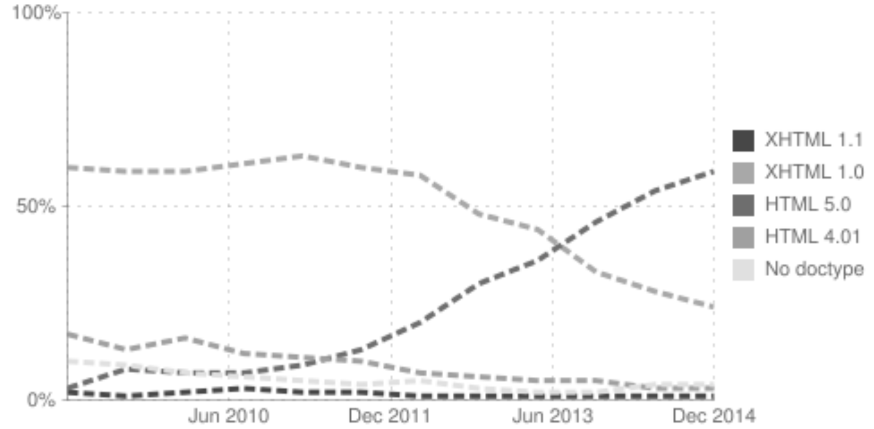


Fig. 1. HTML and XHTML language specifications used by websites.

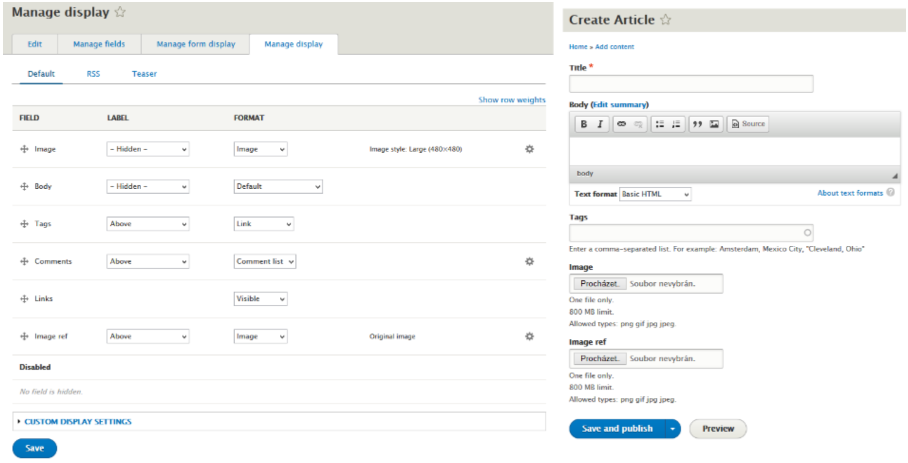
HTML5 specification covers specification of abstract language as well as a set of APIs. The syntax is defined as two types – *HTML 5* based on HTML and XHTML 5 based on XML (XHTML). The HTML syntax comes from the original 4.01 with almost all rules and is recommended for developers. HTML5 is not – by the syntax – an application of some general language. The specification covers complete language definition and does not appeal to any other ones.

Main changes in HTML5 in terms of information content comprise a set of applicable elements and attributes. Some of them are now obsolete including elements *font, listing, strike, big, center*, or attributes *name, summary*, etc. On the other side, there are many new elements bringing new possibilities for semantics.

The future development version – HTML 5.1 (W3 CONSORCIUM, 2015) currently points out only unimportant small changes and improvements. No elements or attributes are removed. In terms of information content, only small improvements on semantics have been recently expected. The specification is currently in Working Draft phase, so changes are expected.

2.4 Content Management Systems and Information Content

Content Management Systems (CMS) store content differently. However, WYSIWYG editors are utilized by most of them for the main part of content. Then it can be extended by another attachments. The WYSIWYG editors can usually add these attachments as well. However, working with such content is limited. Since it is part of the main content, automatic separating is difficult and can lead to inconsistency of the source code. Some CMS solves the problem by composition of blocks of content. E.g. Drupal can add so called fields to post types (nodes). Each node type has a given structure (template), which cannot be changed for node instances. An example is shown in Figure 2.



FIELD	LABEL	FORMAT		
Image	- Hidden -	Image	Image style: Large (480×480)	⚙
Body	- Hidden -	Default		
Tags	Above	Link		
Comments	Above	Comment list		⚙
Links		Visible		
Image ref	Above	Image	Original image	⚙

Disabled

No field is hidden.

Text format Basic HTML

Enter a comma-separated list. For example: Amsterdam, Mexico City, "Cleveland, Ohio"

Image

Procházet. Soubor nevybrán.

One file only.
800 MB limit.
Allowed types: png gif jpg jpeg.

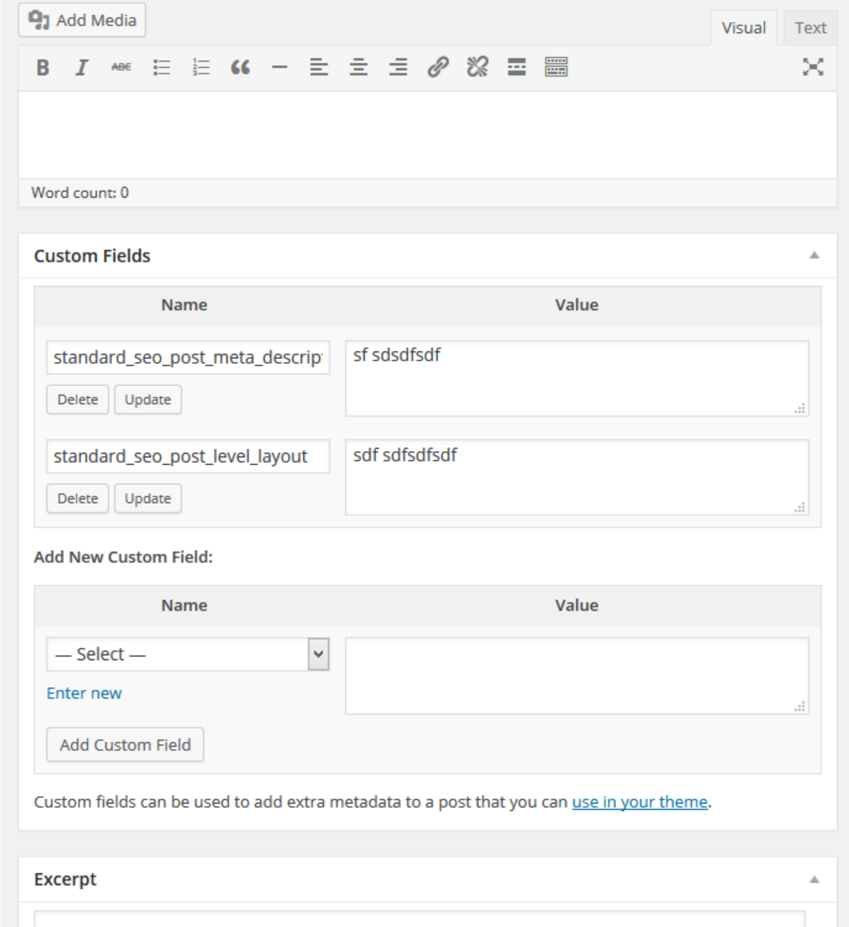
Image ref

Procházet. Soubor nevybrán.

One file only.
800 MB limit.
Allowed types: png gif jpg jpeg.

Fig. 2. Example of creation of node in Drupal CMS. The left side shows defining the structure template. The right side shows authoring of content instance (node)

A change of the template affects all existing node instances (e.g. articles). It can even lead to a data loss. Each field is in a separate table or a table attribute (based on configuration). Other CMS approach these issues in a similar way, an example of WordPress is shown in Figure 3. Content fragmentation dependent on the configuration of certain CMS leads to a limited portability of content. Problems can appear when upgrading to a newer version of CMS, change to a different one or even archiving.



Word count: 0

Custom Fields

Name	Value
standard_seo_post_meta_descrip	sf sdsdfsdf
standard_seo_post_level_layout	sdf sdfsdfsdf

Add New Custom Field:

Name	Value
— Select —	

Custom fields can be used to add extra metadata to a post that you can use in your theme.

Excerpt

Fig. 3. HTML and XHTML language specifications used by websites

2.5 Long term research and investigations

The Department of Information Technologies in cooperation with university newspapers (online magazine) has accomplished several observations and experiments with content authors. The department also operates portal Agris.cz, which is one of the most frequently visited information sources of the agrarian sector in the Czech Republic. Several interviews were run as well.

3 Results

The analysis of HTML in previous chapter was used to determine, how the language specifications changed in terms of information content. Any essential changes were not observed and are not expected for the future.

The analysis of content management systems and other researches suggests the creation of content by composition of separate blocks as a right approach. However, the current solutions are limited due to fixed structure of block. Instances of content are lack of independent ordering of these blocks. This approach also supports the requirements for connection with mobile devices.

Long term research and the cooperation with university newspapers and Agris.cz portal suggests that users have problems with editing advanced content elements. More advanced features of editors can be difficult, especially when it needs some re-edition. The key problem areas can be classified as follows:

- Tables containing pictures
- Floated objects (elements)
- Insertion of automated content parts (from CMS)
- Galleries
- Movements of parts of content inside editor (tables etc.)

Other research results suggest to offer predefined blocks of content to content authors. These blocks are easier for editing and take advantage of the unified final presentation on a website.

4 Conclusion

The results of this study show several conclusions for the further research and project process. The HTML5 can be used for the carriage of main information content parts. In terms of storage, the study suggests to extend it by division into blocks, which can be processed separately. This approach should support the ordering of the blocks independently for each content instance.

These findings suggest that HTML5 language needs to be wrapped in some abstract envelope, which supports the mentioned features. There is a need to define interface for communication and cooperation with content management systems or other applications. Further research should focus on determining the exact form or markup language to be utilized within the target methodology. Using JSON or XML in combination with HTML5 suggests itself to be explored. The following conclusions can be drawn from the presented study:

- The HTML5 language specification is not expected to essentially change

- The language needs to be wrapped in an abstract envelope to support several requirements:
    - Communication with CMS or other applications
    - Allowing support to communicate with mobile devices
    - Universal approach and portability of the content
    - Support for metadata description (e.g. OpenGraph, AGROVOC), correct semantics, sharing (e.g. OAI-PMH), Internet of Things
- Further research should focus on:
    - Useful and user friendly UI (User Interface)
    - Easy and fast content authoring from any device
    - User eXperience measurements
    - Minimizing of input errors in terms of valid and consistent presented output on websites
- Storing of structured data independently on the look of final output
    - Different devices can have different demands on UI

Acknowledgments. The results and knowledge included herein have been obtained owing to support from the following institutional grants.
Internal grant agency of the Faculty of Economics and Management, Czech University of Life Sciences in Prague, grant no. 20151038, “Methods for creation and update of information content in WWW environment“.

References

1. BROWN, Mandy. Writing and Editing in the Browser. The Journal of Electronic Publishing. 2014-02-01, vol. 17, issue 1, s. -. DOI: 10.3998/3336451.0017.111.
2. DAS, S., M. GOETZ, L. GIRARD a T. CLARK. Scientific publications on web 3.0. In: ELPUB 2009 - Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies - Proceedings of the 13th International Conference on Electronic Publishing. Milan, Italy, 2009
3. HTML Version Statistics. PowerMapper [online]. [cit. 2015-05-18]. Available at: http://try.powermapper.com/Stats/HtmlVersions
4. RAGGETT, Dave. Raggett on HTML 4. 2nd ed. Reading, Mass.: Addison-Wesley, c1998, xv, 437 p. ISBN 02-011-7805-2.
5. SPIESSER, J. and KITCHEN, L., 2004. Optimization of HTML automatically generated by WYSIWYG programs, Thirteenth International World Wide Web Conference Proceedings, WWW2004 2004, pp. 355-364.
6. ŠIMEK, P. – STOČES, M. – VANĚK, J. – JAROLÍMEK, J. – MASNER, J. – HRBEK, I. Using of Automatic Metadata Providing. AGRIS on-line Papers in Economics and Informatics, 2013, 5(4), 189-197. ISSN: 1804-1930.

7. ŠIMEK, Pavel, Michal STOČES and Jiří VANĚK. Mobile Access to Information in the Agrarian Sector. AGRIS on-line: Papers in Economics and Informatics. 2014, 6(2): 89-96. ISSN 1804-1930

8. Total number of Websites. Internet Live Stats [online]. [cit. 2015-02-02]. Available at: http://www.internetlivestats.com/total-number-of-websites/

9. W3 CONSORCIUM. HTML 4.01 Specification [online]. W3C Recommendation 24 December 1999. 1999 [cit. 3.5.2015]. Available at: http://www.w3.org/TR/html4/

10. W3 CONSORCIUM. XHTML™ Modularization 1.1 - Second Edition [online]. W3C Recommendation 29 July 2010. 2010 [cit. 3.5.2015]. Available at: http://www.w3.org/TR/xhtml-modularization/

11. W3 CONSORCIUM. XHTML™ 1.1 - Module-based XHTML - Second Edition [online]. W3C Recommendation 23 November 2010. 2010 [cit. 3.5.2015]. Available at: http://www.w3.org/TR/xhtml11/

12. W3 CONSORCIUM. HTML5: A vocabulary and associated APIs for HTML and XHTML [online]. W3C Recommendation 28 October 2014. 2014 [cit. 3.5.2015]. Available at: http://www.w3.org/TR/html5/

13. W3 CONSORCIUM. HTML 5.1 [online]. W3C Working Draft 17 April 2015. 2015 [cit. 3.5.2015]. Available at: http://www.w3.org/TR/html51/