WWW.HR: A tool for editing Dublin Core metadata information in the Croatian Web space

Gordana Stojšić, Igor Ljubi
University of Zagreb
Department of Telecommunications
Faculty of Electrical Engineering and Computing
Unska 3, HR-10000 Zagreb

E-mail: igor@tel.fer.hr, gogs7@net.hr

Abstract

Using metadata in Web pages has been recognized as one of the ways to make Web search more efficient. In the area of metadata standardization, the Dublin Core metadata standard has been developed for describing a wide range of networked resources. The Croatian Academic and Research Network recommend the use of Dublin Core metadata in order to improve the searching of the Croatian Web space.

Metadata can be added to a HTML document manually, by editing the source, but this approach, aside from being tedious and time-consuming, may also introduce unintentional errors in typing, HTML syntax, etc. In order to make adding metadata easier, a tool for editing Dublin Core metadata was created. This tool allows introducing new metadata, as well as viewing and modifying existing metadata in an HTML file. Finally, the tool also automatically extracts potential keywords from the text, thus making the selection of keywords easier for the Web administrator. The keyword extracting algorithm is based on the frequency of words in the text as well as their position and formatting emphasis in the text.

This tool was tested on the content of the WWW.HR information on Croatia covering the topics history, economy, culture, tourism, nature etc.

Introduction

For more than eight years, WWW.HR has been providing the basic information about the Republic of Croatia to the surfers worldwide. It's two award winning informational services, Short info on Croatia and the Directory of Croatian Web sites have attracted millions of hits from surfers who wanted to get some knowledge about Croatia. Great user's response, as well as constant growth of visitors was enormous encouragement for upgrading our services. One of more recent upgrades was last year's addition of Dublin Core metadata to all our pages on sites that are providing short info on Croatia. Following that effort, and to encourage our users to put metadata on their sites, we have developed a tool for editing Dublin Core metadata.

This paper is organized as follows: In Section 1 Dublin Core standard and CARNet's recommendation are presented, Section 2 offers comparison to other available metadata generators, while Section 3 deals with the requests for the tool during the designing phase. Section 4 puts theory into practice, and shows an example using data available in WWW.HR directory. Section 5 states what would be the next steps and concludes the paper.

1. How to make searching the Web easier?

It is said that every information know to mankind can be found on the Internet. And yet, there is no-one who haven't encountered big problems when he/she tried to find information, but hasn't know the exact location on the Web to find it. Users usually try two solutions to satisfy their thirst for information. They either try to find the desired information using some directory of the Web sites, or they enter search pattern into one of the Internet search engines. But the results from the directory may prove to be inadequate, especially if user wants to reach a bit more specific information. On the other hand, conducting a search can return a few millions results, making it almost impossible to find what user really wants to read.

One of the reasons of this chaos is the fact that documents on the net are rather badly described and indexed. The possible solution to this very important issue was to somehow accurately describe these documents. This is the point of inserting the metadata. Metadata is "data about the data". Documents described with metadata are much more searchable and easier to index.

As the result of the need to better describe Internet's documents, and as the result to make them easier to find, a group of librarians, content experts and IT specialists made a set of describing elements, known as Dublin Core Metadata Element Set (DCMES). DCMES has 15 metadata elements, as shown in Table 1.

Content		Intellectual property	Instantion
Title	Source	Creator	Date
Subject	Relation	Publisher	Language
Description	Coverage	Contributor	Format
Type		Rights	Indentifier

Table 1 – Dublin Core Metadata Element Set

Based on the experiences gained through WWW.HR and Croatian Search Service (CROSS) projects, CARNet has issued a recommendation CDA0027 [12], for enhancing a searchability of the Croatian Web space by using the metadata. For further details about how we implemented this recommendation to WWW.HR site, an interested reader is referred to [5].

2. Available Metadata Generators – State of the Art

Before trying to implement our solution, we have investigated the possibilities of the currently available metadata generators. Four free tools available on the Web have been selected for further analysis. Those tools are:

- *Nordic DC metadata creator* (http://www.lub.lu.se/cgi-bin/nmdc.pl) [6]
- *DC-dot* (http://www.ukoln.ac.uk/metadata/dcdot/) [7]
- *Reggie* (http://metadata.net/dstc/) [8]
- *Klarity* (http://www.klarity.com.au/) [9].

Following tool's characteristic was studied and used for comparison purposes:

- form that has to be filled in order to generate metadata
- extracting metadata from HTML
- can convert DC metadata in other types of metadata
- editing metadata
- multi language support

Results of the testing on free tools are given in Table 2.

	Form metadata	Metadata extraction from HTML	Converting schemes	editing	Multi language support
Nordic DC metadata creator	+		+	+	
DC-dot		+	+	+	
Reggie		+	+	+	+
Klarity		+			

Table 2 – Results of testing free metadata generators

It is clear that neither of free tools complies with all five characteristics that were studied. Only one tool supports multiple languages, and three out of four are able to extract metadata from HTML, convert DC metadata to other metadata schemes and to edit the metadata that they generate.

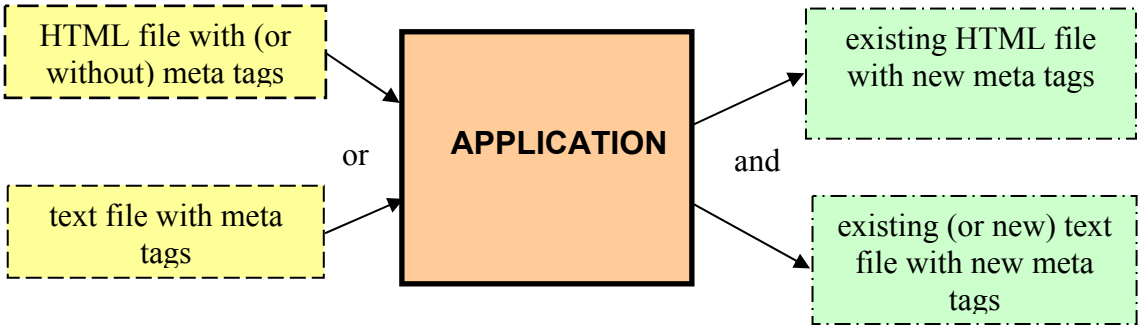
Beside the free tools, there are also a few commercially available solutions. One of them is *Metabrowser*. It's a Web browser that is simultaneously showing both Web page and the metadata that describes it. *Metabrowser* can create and edit metadata. It supports various metadata schemes (DC, GILS, AGLS, EdNA), which can be loaded either from hard disk or from the Internet. *Metabrowser* is available at http://metabrowser.spirit.net.au/ [10].

Another commercial tool for editing metadata is Gen – Dublin Core Edition. This tool has graphical user interface that is used for creating the metadata. It also enables the creation of other meta tags important for search engines. The tool can automatically add existing metadata, generate keywords, and can also show how the search engines will see that Web site. This tool is available at http://bridges.state.mn.us/taggen.html [11].

3. Requests on the tool

The main task of the tool is to allow editing and viewing Dublin Core metadata from a HTML document. The created DC metadata should be inserted in the <HEAD> tags of a HTML document. Croatian language has specific characters, which are typical for the South Slavic languages and they can't be found in all character encodings. This tool should be able to handle these typical characters. In 'CDA0027' document CARNet recommends use of Dublin Core metadata in order to improve searching of the Croatian web space. Metadata should be created in accordance with this document and Dublin Core recommendations. One of the future users request was the automatic keyword generation. Another request was creation of a text file, which contains only metadata. Every edited HTML file should have a corresponding text file. This text file could be edited with a text editor and changes made in the text file could be applied on the HTML file using this tool.

The application was created according to these requests. Picture 1 shows the possible inputs of the application and the outputs. Inputs are in yellow rectangles (dashed line) and the outputs are in green rectangles (dash-dot line). The application can edit metadata from a HTML file or from a text file, but its outputs are always a HTML and a text file with new metadata.



Picture 1 – Inputs and outputs of the application

4. Dublin Core Metadata Generator for use in Croatian Web space

4.1. Editing the metadata with the tool

The tools interface allows easy editing and viewing DC metadata.
This tool supports these two encodings:

- ISO-8859-2 (Latin 2)
- MS-CP1250 (MS Latin II)

Which encoding will program use depends on the command line parameter.
If DC.Subject, DC.Title and DC.Date.Modified meta tags don't exist in a HTML document the program computes values for these metadata. In this way, the program facilitates creating metadata for a document. The content of <TITLE> tags is suggested as DC.Title meta tag value and the current date is suggested as DC.Date.Modified meta tag value. The program

extracts keywords from the document in order to compute the DC.Subject meta tag value.

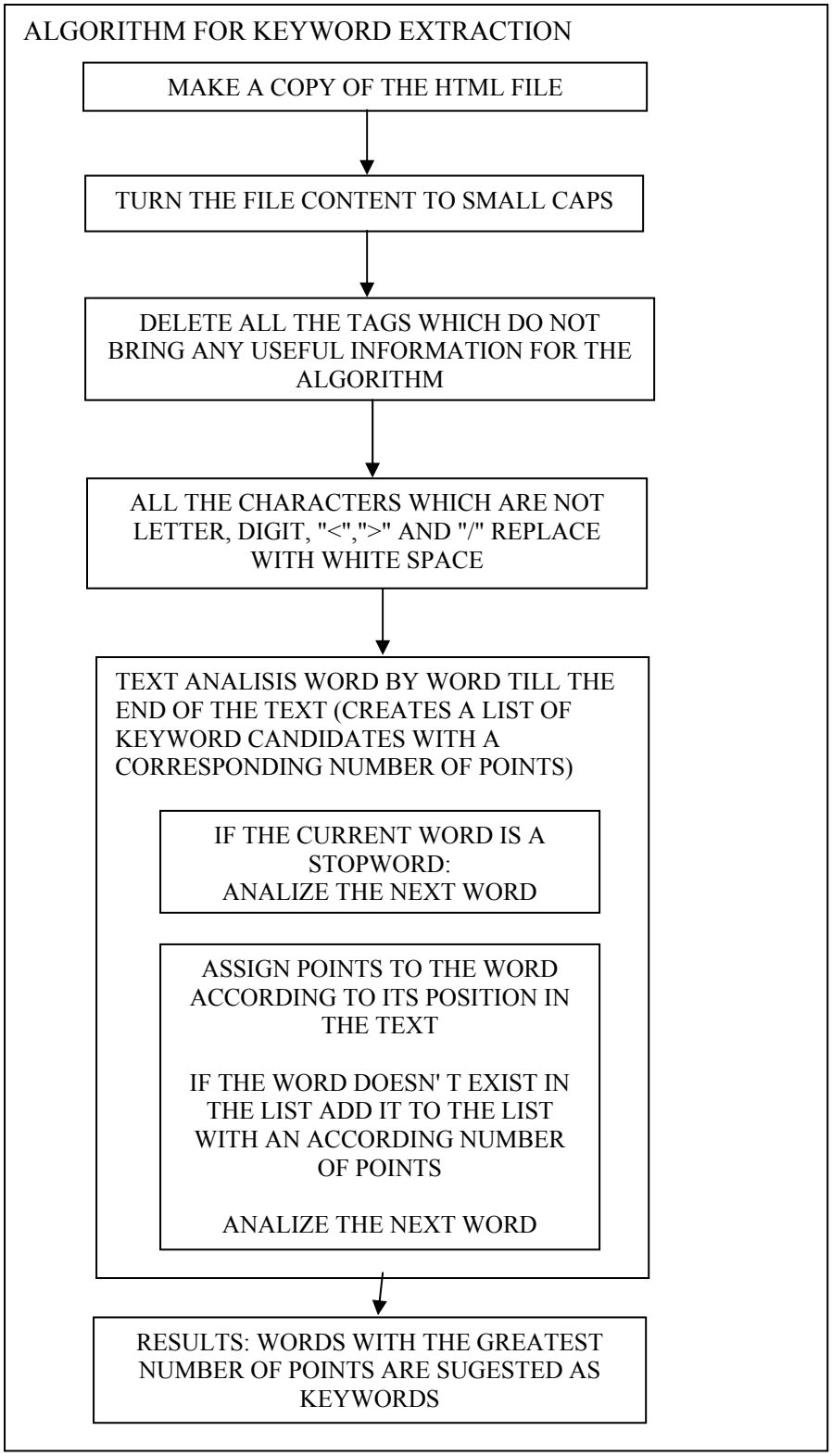
Automatic keywords extraction algorithm is based on the frequency of word occurrences and on the location in which they occur. More times the word appears in the text, more probably it will be extracted as a keyword. In a similar way when a word appears in a 'title' tag (title tags are: <TITLE>, <H1>, <H2> and <H3> tag) it is considered 'more relevant' than a word that appears in the rest of the text. Every appearance of a word brings a corresponding number of points to the word according to these two criteria. Here is the number of points assigned to a word for its appearance in the text:

- Appearance in a <TITLE> tag: 23 points
- Appearance in a <H1> tag: 15 points
- Appearance in a <H2> tag: 10 points
- Appearance in a <H3> tag: 6 points
- Appearance it the rest of the text: 2 points

Words with the greatest number of points are suggested as keywords.

The text processing has the following phases:

- The document letters are transferred to small caps
- All the characters, which are not letters, digits, or characters that are marking beginning or the end of a tag are replaced with white spaces
- All the tags that are not 'title' tags are replaced with white spaces
- Stop words aren't considered candidates for keywords. Stop words are adverbs, prepositions, junctions, and other words like 'com', ' is', 'are', 'html' which often appear in the text but don't bring useful information about it's content.



Picture 2 – Keywords extraction algorithm

Keyword extraction algorithm is shown in Picture 2.

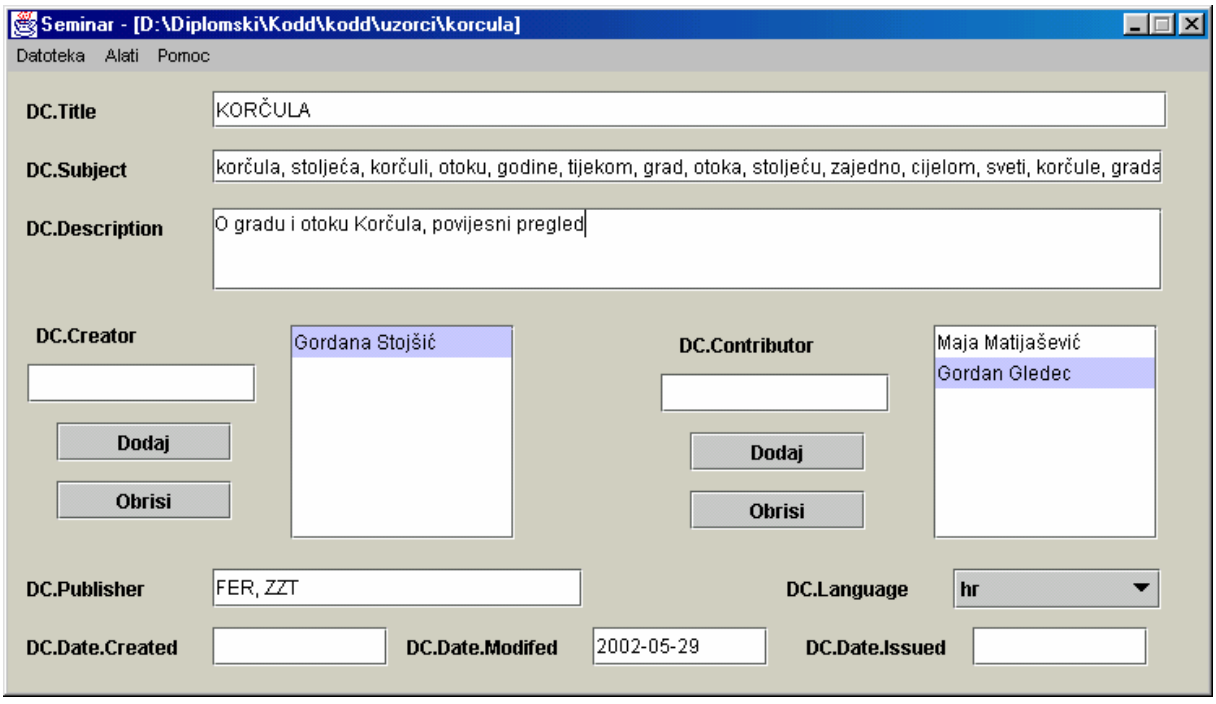
This tool has following functions, which facilitate editing metadata:

- 'Adding the template' (The template contains the most frequently added metadata values (values for DC.Creator, DC.Contributor and DC.Publisher). When the user chooses this option, the template is added to the interface.)
- 'Computing keywords' (This option recomputes the keywords for the document)
- 'Rebuilding the list of stop words' (By choosing this option user can rebuild the list of stop words and recomputes keywords.)
- 'Adding keywords' (If a user wants to add the extracted keywords to existing ones it can be done by choosing this option.)

After editing metadata a user can choose between 'Save' and 'Save as...' options. 'Save' option replaces old metadata with new ones and creates a text file with metadata if it doesn't exist. If the option 'Save as...' is chosen the program creates a new HTML and a new text file with metadata.

4.2. Application on the content of the WWW.HR service

Picture 3 shows the tool's interface. The opened HTML document didn't have DC metadata. The program computed values for DC.Title, DC.Subject and DC.Date.Modified. On the picture we can also see the result of the user's selection 'Adding the template'. Values for DC.Creator, DC.Contributor and DC.Publisher are added to the interface.



Picture 3 – Tool's interface

4.3. Created metadata

The following metadata (text file with metadata) were created with the tool and correspond to the metadata values shown on the Picture 3.

```
<META name="DC.Title" content="KORČULA">
<META name="DC.Description" content="O gradu i otoku Korčula, povijesni pregled">
<META name="DC.Subject" content="korčula, stoljeća, korčuli, otoku, godine, tijekom, grad, otoka, stoljeću, zajedno, cijelom, sveti, korčule, grada, pod, vlast, živjeli, tom, ">
<META name="DC.Creator" content="Gordana Stojšić">
<META name="DC.Publisher" content="FER, ZZT">
<META name="DC.Language" content="hr">
<META name="DC.Date.Modified" content="2002-05-29">
<META name="DC.Contributor" content="Maja Matijašević">
<META name="DC.Contributor" content="Gordan Gledec">
```

5. Conclusions

This paper presents the work conducted for the diploma thesis. The objective of the study was to create a usable tool for generating and editing metadata information. Dublin Core Metadata Element Set, as a de facto standard, was the obvious choice for metadata scheme to be used with the tool. By studying other similar tools, architecture for this tool was designed. Because this tool is to be used within Croatian Web space, some additional requirements were made (e.g. Croatian specific letter had to be properly found). That problem was solved using proper encodings. Keywords used for the creation of the metadata were extracted using a special keyword extraction algorithm. A GUI was created to help users to fill form for adding and editing metadata.

This tool was tested on both of the WWW.HR informational services. Testing of the tool on those pages proved to be easy and efficient in generating and editing metadata. Metadata can easily be put into HTML documents.

Hopefully, this tool will be used to generate sufficient amount of metadata to enable better searchability of Croatian Web space.

References

[1] Gledec G, Jurić J, Matijašević M, Mikuc M: WWW.HR - Experiences with Web-server Development and Maintenance, Proceedings of the XXII. International Convention MIPRO 99, p. 83-86, Opatija, svibanj 1999.

[2] Ljubi I, Gledec G: *WWW.HR – An entry point to the Croatian Cyberspace*, CARNet Users' Conference, CUC 2000, Zagreb, rujan 2000.

[3] Ljubi I, Gledec G, Matijašević M: *WWW.HR – The Rise of a National Web Portal*, Proceedings of the International Symposium on Telecommunications VITEL 2000, p. D39 – D42, Ljubljana, Slovenija, listopad 2000.

[4] Ljubi I, Gledec G, Matijašević M: *WWW.HR directory: Adding value by use of metadata*, Libraries in the digital age – LIDA 2001, Dubrovnik, svibanj 2001.

[5] Ljubi I, Gledec G: *WWW.HR: First metadata-enabled service in Croatian Webspace*, CARNet Users' Conference, CUC 2001, Zagreb, rujan 2001.

[6] Nordic DC metadata creator (including URN generator), Preben Hansen, 1999, http://www.lub.lu.se/cgi-bin/nmdc.pl

[7] DC-dot (UKOLN), Andy Powell, 2000, http://www.ukoln.ac.uk/metadata/dcdot/

[8] Reggie (DSTC - Australia), DSTC RDU, 1998, http://metadata.net/dstc/

[9] Klarity, tSA Consulting Pty Ltd, 2000, http://www.klarity.com.au/

[10] Metabrowser, Spirit Consulting, 2000, http://metabrowser.spirit.net.au/

[11] TagGen – Dublin Core Edition, Olson, Robert, 2001, http://bridges.state.mn.us/taggen.html

[12] Maja Matijašević, Hrvoje Stipetić, Kako poboljšati pretraživost hrvatskog World Wide Web prostora uporabom metapodataka, 2001, ftp://ftp.carnet.hr/pub/CARNet/docs/advisories/CDA0027.doc

Biography:

GORDANA STOJŠIĆ received her Diploma in Electrical Engineering from the Faculty of Electrical Engineering and Computing, University of Zagreb, in 2002. Her research interests include Internet technologies and Java. This paper presents results from her undergraduate thesis, titled "Introducing metadata in the WWW site content".

IGOR LJUBI received his B. Sc. E.E. from the Faculty of Electrical Engineering and Computing, University of Zagreb, in 1999. He has been working at the Faculty of Electrical Engineering and Computing as an associate assistant since March 1999. His research interests include software engineering, mobile agents and WWW programming. He is involved in a CARNet project WWW.HR – Homepage of the Republic of Croatia since 1999. He is a member of the IEEE, and is actively involved in IEEE Student Branch Zagreb.