XML-based authoring support for the creation and management of multilingual information resources

Dietmar Rösner, Uwe Dürer, Mario Krüger, Sandro Neils

Otto-von-Guericke-Universität Magdeburg
Institut für Wissens- und Sprachverarbeitung
P.O. Box 41 20, D–39016 Magdeburg, Germany
roesner@iws.cs.uni-magdeburg.de

Abstract

An XML based approach to content mangement is presented that relies on annotating information objects with metadata, structural markup and inline semantic tags. To ease the work of content creators an authoring tool has been implemented. Labels used in mark up are given a precise meaning because they are organised in a domain specific ontology.

Keywords: content management, metadata, document structuring, authoring tool, ontology

1 Introduction

The management of a pool of information objects (i.e. texts, pictures, videos etc.) is still a challenge especially when different authors are contributing to this pool. The problems are even more complicated when the information systems based on the pool of information objects should be multipurpose – rearrangeable for different types of usage – and multilingual – available in a number of different languages. Such a situation is e.g. typical for technical documentation that should be different for the technician and service personnel as opposed to the end consumer and that should be available in a multitude of languages. In our project CATCH[1] we are confronted with similar demands for the management of information objects for health information systems for the European citizen as end users. In this paper we present our approach to content management based on the use of XML in combination with a domain specific ontology.

The paper is organised as follows: We first give background information and sketch the objectives of the project CATCH. We then present the XML based approach to content management problems and emphasize the advantages of separating structuring and storage issues from all aspects of presentation and layout. An example text is used to illustrate the different types of XML tags

[1] The work reported has been performed in the project CATCH-II that is funded by the European Commission (DG Information Society) under contract HC 4004.

used in CATCH. This is followed by a detailled description of the functionality of the implemented authoring support tool CEdit. The paper ends with a discussion of related work and a summary.

2 CATCH: health information for citizens – background and objectives

CATCH – an acronym for **C***itizen* **A***dvisory System based on* **T***elematics for* **C***ommunication and* **H***ealth* – is a European project in the Telematics Application Program. Its goal is to create a framework for telematics based health information systems for the general public in the EU countries. The various prototypes developed are evaluated with a variety of groups of citizens from different EU countries. The users differ with respect to language and cultural background, age, education level, health awareness etc.

The outcome of CATCH will not be a (software) product as such but rather an elaborate methodology for the effective design of such systems with respect to issues of e.g. content structuring, efficient use of interactivity, tailoring of presentation to the users needs and preferences and maintaining multilingual information services.

In addition concepts and prototypical software solutions (tools) for authoring, maintaining and updating such information systems will be developed and tested. At best these tools will be usable by content providers without the need for programming skills and other specialised knowledge about computer science and telematics.

2.1 The core issues: authoring support and metadata

Cost and time needed for authoring is still a major limitation for the creation and delivery of multimedia information systems. Therefore the work in the CATCH project concentrates on concepts and steps towards a methodology for supporting the authoring process.

In the CATCH project content is created by medical experts from different European countries. These authors primarily do write texts about medical topics from their field of expertise, but may additionally provide pictures, graphics and videos as well. Due to their relevance in disease prevention the example topics of skin cancer and cardiovascular diseases have been chosen for the project.

In order to maximise the chance for reusability and flexible exploitation of the created information objects we have decided to use a global architecture based on XML that completely separates content creation and storage from all aspects of delivery and layout.

Authors do write their texts and are asked to both annotate them with metadata (encoded in XML, *extensible markup language*, [2], [3])) as well as structure them with XML markup and XML inline tags for major elements of importance in medical texts (e.g. disease, body part, pharmaceutical, ...). To ease this process CEdit, a tool for metadata capture and for XML tagging, has been designed and implemented as part of the CATCH authoring support (cf. below).

Authors then submit their contributions to the central CATCH information pool (or repository). From this database with XML tagged information objects delivery versions of CATCH are created and updated. This will in the future ideally be highly automatised, in the current stage of development human intervention is still necessary.

2.2 Exploiting markup and metadata

It is an additional effort to annotate documents with metadata and logical markup. On the other hand their availability can play a crucial role for the management, update, reusability and exploitation of the information resources. Some examples from CATCH illustrating the added value:

- Textual resources (e.g. texts about diseases, prevention, lifestyle etc) have to be kept 'in parallel' in all the European languages covered. The project decided to use the English text version as master version. Metadata will help to manage the workflow of keeping versions of the textual resources in different languages consistent, e.g. when there is a change in the master version of a text an update of the parallel texts can automatically be stipulated.

- The CATCH system is available via the Internet but specialised versions tailored to special environments (e.g. kiosks in hospitals) will have to be easily configurable from the total information pool. This can be organised by extracting only those information objects from the information pool that deal with topics relevant for the specialised version.

- A tool to exploit the inline tags together with a domain ontology will help in automatically linking new textual resources with already existing ones (e.g. link the mention of a disease in a text with the definition text of this disease, the mention of a pharmaceutical with a descriptive page for this pharmaceutical, ...).

3 A closer look: tagging in CATCH

The XML tags used in CATCH information resources fall into the following broad categories:

- tags capturing metadata that are of a 'bibliographic' nature,

- tags for delimiting structural units in the texts,

- inline tags for 'typing' natural language terms in the texts.

These categories are best explained through concrete examples. We will use the following excerpt from an example text about skin cancer for this purpose ('...' indicate ellipsis).

```
<?xml version="1.0" ?>
<CATCH-INFO-ELEMENT>
<META authors="Dr. Schramm, Luckert" supervisor="Prof. Gollnick"
copyright="UDV, 1999"/>
<META translated-by="DR" translation-date="March-15-99"
time-to-translate="50min"/>
...
<Body>
<Question-Answer-Pair>
<Question>
How to perform <PROC-Diagnostic>self examination and self diagnosis!
</PROC-Diagnostic>
</Question>
<Answer>
In prevention (of skin cancer) we distinguish between <DIS-Prevention>
primary prevention</DIS-Prevention> through information and early
detection as <DIS-Prevention>secondary prevention</DIS-Prevention>.
<Appeal>
There is no doubt: In addition to primary prevention early detection
plays the most significant role in the fight against cancer.
Don't forget: you are the most important factor in early detection!
</Appeal>
It is a big advantage that the <Organ System>skin</Organ System>
-- in contrast to many other organs  -- is visible and can be examined
without technical devises and <PROC-Diagnostic>without invasive
examination methods</PROC-Diagnostic>.  We thus have the basis for an
examination method applicable by everybody.

For all those <DIS-Etiology>malignant diseases of the skin</DIS-Etiology>
that develops visibly regular self examination offers a big chance
to detect <Disease>cancer</Disease> already in an early stage.
...
```

3.1 Bibliographic metadata:

Bibliographic metadata have the purpose to capture information about the process of creating, translating and updating information resources. We take a set of tags inspired by the Dublin Core ([6]) as baseline but extend it with tags needed for our purpose.
In the example text the metadata about the authors, the supervisor, the copyright holder and the dates about the translation are bibliographic metadata.

3.2 Structural units:

In order to maximize flexibility and reusability of the information resources CATCH exploits the advantages of an architecture based on the complete separation between logical document structures and all aspects of layout and delivery formats.

The repository of CATCH information resources (the CATCH information pool) does contain textual information units that contain logical tags only. The mapping from logical document structures to delivery formats or the on line dynamic creation of layout is organised as a separate process.

Structural tags allow the authors to give interpretable clues to the processes that operate on the information units what parts there are and what their respective purpose is.

For the example text we suggest to interpret it as a structural unit named `<QUESTION-ANSWER-PAIR>` comprised of the two elements named `<QUESTION>` and `<ANSWER>`. This is based on the observation that the head sentence (i.e. `How to perform ...`) serves to pose a (rhetorical) question that is answered by the rest of the text. Please note that such an analysis prefers underlying speech acts over surface syntactic appearance. The head sentence could as well have been paraphrased as *'How can I perform self examination and self diagnosis?'* or as a nonsentential head *'Self examination and self diagnosis'*.

The `<ANSWER>` subunit captures the rest of the information element.

Given this structural organisation the layout of the `<QUESTION>` will underline its function as a title of the information unit. In the example case the unit is part of a series of `<QUESTION-ANSWER-PAIR>`s. So all `<QUESTION>`s might be used as index to their respective `<ANSWER>` texts and e.g. given in a dynamically created overview page.

In the example text there are other units marked according to their function:

- There is an explicit appeal:

```
<APPEAL>There is no doubt: In addition to primary prevention early
detection plays the most significant role in the fight against cancer.
Dont forget: You are the most important factor in early detection!
</APPEAL>
```

With an appeal an author is trying to influence the readers general attitudes and motivations towards a certain topic (in the above case: regular self examination of the skin).

- There is a recommendation:

```
<RECOMMENDATION>The examination should be carried out
<DETAIL>with a completely nude body</DETAIL>.</RECOMMENDATION>
```

Recommendations are related to aspects of the readers behaviour. The author is suggesting to perform certain actions or to perform them in a specific manner. In a negative version they may recommend to avoid certain actions.

Again the layout may treat these structural units differently and highlight there function (e.g. through different color, font, highlighting, etc.).

A more complete list of structural units is under development together with the authors. For the start we suggested to include as well tags for e.g.:

– `<DEFINITION>`

A `<DEFINITION>` comprises the term to be defined (the `<DEFINIENDUM>`) and the defining text (`<DEFINIENS>`).

- <WARNING>
A <WARNING> has the purpose to make the reader aware (or again aware) of a possible risk or danger.

3.3 Medical terminology:

The ontology used in the CATCH authoring environment has the primary purpose to help the authors in annotating their textual contributions to the CATCH pool of information resources with standardised categories.

The core of the ontology is an organisation of medical terms. It is primarily organised from the perspective of layperson but takes as well a professional perspective into account. In addition it is intended to take emerging quasi standards like the UMLS (unified medical language system, [7]) into account.

Some terms will be used as metadata to characterise the content of the information unit as a whole, others are used to tag terms inline in the text to make their (semantic) type explicit and available for further processing.

4 CATCH: How authors are supported

A reoccurring question for this approach is: What will authors have to do? How will they be supported?

In the simplest case, authors could simply write their texts in their favorite text editor and submit it as plain ASCII file without any tags to the administrators of the CATCH information pool. In order to have the advantages of structuring and inline tagging the text would then need a posteriori tagging.

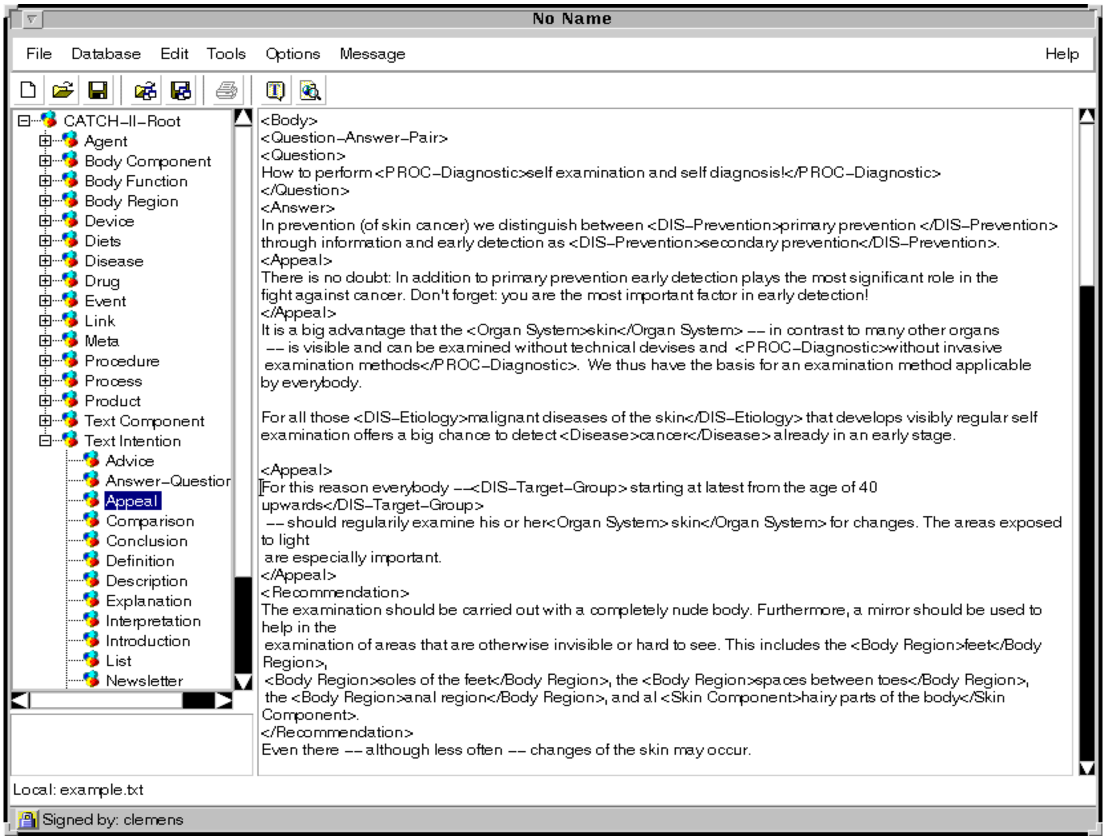
We work with the hypothesis that authors will prefer to have the closer control over their products offered by the XML tags and therefore prefer to do the tagging integrated in their text production and editing process. To allow this in a user friendly manner is the purpose of the CATCH authoring support tools.

4.1 CEdit - XML based authoring support

The Cedit application is a special software tool which has been developed for the authors of health related content. It is available for all authors from the four countries participating in the CATCH-II project who contribute medical or health related information to the CATCH-II system. That means, CEdit is available in different languages, but with a uniform user interface.

The Cedit application offers the following functionality:

- Selection of the language for user interface, help texts, etc.
- Creation (writing) of new texts and saving of these texts – in plain ASCII format or annotated as XML file – on the author's local hard disk.
- Annotating a text with metadata
- Insertion of structural markup into a text
- Insertion of inline semantic tags based on the CATCH-II ontology



```
<Body>
<Question–Answer–Pair>
<Question>
How to perform <PROC–Diagnostic>self examination and self diagnosis!</PROC–Diagnostic>
</Question>
<Answer>
In prevention (of skin cancer) we distinguish between <DIS–Prevention>primary prevention </DIS–Prevention>
through information and early detection as <DIS–Prevention>secondary prevention</DIS–Prevention>.
<Appeal>
There is no doubt: In addition to primary prevention early detection plays the most significant role in the
fight against cancer. Don't forget: you are the most important factor in early detection!
</Appeal>
It is a big advantage that the <Organ System>skin</Organ System> –– in contrast to many other organs
 –– is visible and can be examined without technical devises and  <PROC–Diagnostic>without invasive
 examination methods</PROC–Diagnostic>.  We thus have the basis for an examination method applicable
by everybody.

For all those <DIS–Etiology>malignant diseases of the skin</DIS–Etiology> that develops visibly regular self
examination offers a big chance to detect <Disease>cancer</Disease> already in an early stage.

<Appeal>
For this reason everybody ––<DIS–Target–Group> starting at latest from the age of 40
upwards</DIS–Target–Group>
 –– should regularily examine his or her<Organ System> skin</Organ System> for changes. The areas exposed
to light
 are especially important.
</Appeal>
<Recommendation>
The examination should be carried out with a completely nude body. Furthermore, a mirror should be used to
help in the
 examination of areas that are otherwise invisible or hard to see. This includes the <Body Region>feet</Body
Region>,
 <Body Region>soles of the feet</Body Region>, the <Body Region>spaces between toes</Body Region>,
 the <Body Region>anal region</Body Region>, and al <Skin Component>hairy parts of the body</Skin
Component>.
</Recommendation>
Even there –– although less often –– changes of the skin may occur.
```

Figure 1: A screenshot of CEdit (with part of the CATCH ontology in the left column)

- Search for texts already existing in the CATCH-II information pool
- Submission of annotated texts as new contribution to the CATCH-II information pool

With the help of this tool, the authors have the possibility to generate medical texts, to specify the content of texts using metadata, and to include medical as well as structural XML tags into the text body.

The CATCH-II team has developed this tool with the help of the programming language Java. CEdit is implemented as a Java Applet (i.e. a sequence of Java commands that is embedded in a webpage) which rests on the server side and runs in the browser window of the end user (the author). In this way, the application can be updated very easily by an exchange of the appropriate Java applet on the server. That means, the authors will neither be involved in any installation procedure nor in the management of this utility.

4.1.1 CEDIT in detail

For a better understanding of the working of this application, it is useful to follow the workflow during the authoring process with CEdit:

At the beginning of each session, the author starts his/her browser application (e.g. Microsoft's Internet Explorer or Netscape's Navigator) and accesses the URL of CEdit. When the start menu of the Java Applet appears in the browser window, the authors are asked to identify themselves with user name and password. In this way, it can be assured that only registered authors have full access to the database. Furthermore, the information about the user can be associated with the current text.

After the initial registration and – optionally – language choice, CEdit loads the list of labels for metadata and XML tags as well as an index of topics available already in the information pool.

With the help of the editor tool, the authors are capable to create new texts or to modify already existing ones. For that purpose, the editor tool will be equipped (in a future version) with import filters for the most widely distributed formats of text applications (ASCII, Winword, ...), as well as for HTML and XML files.

Autors can retrieve a list of all titles of texts already existing for a certain topic. If the author wants to modify one of these texts instead of creating a new one, he can load it from the information pool.

At the end of the preparatory phase, the author can open a context menu and input (or modify) metainformation. Identifiers for bibliographic as well as content oriented metadata are available from a rollup menu.

In the next step, the inline tagging of text elements follows. For that purpose, the appropriate target words or word groups must be selected first, then it is possible to select one of the CATCH-II specific XML tags. At the end of the tagging procedure, the authors have the possibility to verify the syntax. Error messages appear if the authors break the rules that are embedded in the CATCH-II DTD.

Moreover, the editor tool has an INSERT menu. This menu contains functions for the embedding of hyperlinks and multimedia objects. Here, the author can specify the location of the target object.

The final result of the editing process can be saved locally or submitted to the central information pool. A preview function helps the authors to judge the result of the working session.

4.1.2 Delivery issues

For the delivery of XML documents in browsers there are two major options:

– Option 1: Using stylesheets a browser can display XML documents directly. There is fast development in the field of XML aware browsers but for now we found it wise to experiment as well with option 2.

– Option 2: When configuring CATCH runtime versions (either for Internet or offline versions) the texts and other resources from the CATCH pool are mapped into a collection of linked HTML pages for display in any HTML-based browser.

The editor tool creates only XML source code files. However, for the display of the final version of a webpage, external style sheet files (XSL, CSS2) and DTD files (document type definitions) are needed. Such files can also be placed on the hard disk of the author's computer. In this way, it is possible to use an internet browser for an offline preview of the text.

5 Discussion

5.1 Relation to other work

It is nearly a common place now that the wealth of information available in the WWW can only be exploited when information units are annotated with metadata about their content ([1]). It is as well widely agreed that a shared ontology as a common conceptualisation is needed for the interpretation of metadata and semantic annotations.
The projects Ontobroker [4] and its follow up On2broker [5] for example employ an ontology based approach for providing query interfaces to existing web pages. HTML pages are annotated with concepts from an ontology either manually or – when the pages' structure is stable enough – with the help of so called wrapper programs. Our approach differs in that we have the chance to integrate the mark up into the authoring and creation process and thus do not have to deal with the problems of a posteriori tagging of existing pages.

5.2 Ongoing work in CATCH

In our approach the creators of documents are stipulated to provide metadata, to structure their documents logically and to provide semantic typing for relevant terms used in the texts. This extra work of authors is supported by a tool that eases annotation. The suggested tags are organised in a taxonomy and offered in a menu based interface.
The experiences of the content providers, i.e. medical experts, with both the methodology and the supporting tools will be a central topic in the evaluation phase of CATCH that is about to start at the time of this report.
Other issues on the agenda include:

- better exploitation of the metainformation associated with the texts in the CATCH information pool,

- extension of the approach to non-textual information units as well.

5.3 Beyond CATCH

Although developped for the support of authors in CATCH the tools implemented are by no means limited to this application. Their design guarantees that they could just as well be used in other applications with a need to manage structured information resources. Application areas like the management of multilingual technical documentation or other business applications are obvious candidates. The primary change necessary for a transfer to another application would be to (simply) supply an ontology appropriate for the new domain.

System availability: CEdit can be accessed and tested via the URL

`http://catch.cs.uni-magdeburg.de/CEdit/index.html`

with a guest account (user: guest, password: visitor). The authors would welcome feedback from visiting users.

Acknowledgements. We would like to thank the anonymous reviewers for constructive criticism and valuable suggestions.

References

[1] Murtha Baca. *Introduction to Metadata – Pathways to Digital Information.* Getty Information Institute, 1998.

[2] Jon Bosak.
XML, Java, and the Future of the web. *http://sunsite.unc.edu/pub/sun-info/standards/xml/why/xmlapps.htm*, 1996.

[3] Tim Bray, Jean Paoli, and C.M. Sperberg-McQueen. Extensible Markup Language (XML) 1.0. *http://www.w3.org/TR/1998/REC-xml-19980210*, 1998.

[4] Stefan Decker, Michael Erdmann, Dieter Fensel, and Rudi Studer. ONTOBROKER: Ontology based Access to Distributed and Semi-Structured Information. In R. Meersman et al. (eds.), editor, *Semantic Issues in Multimedia Systems.* Kluwer Academic Publisher, Boston, MA, 1999.

[5] Dieter Fensel, Juergen Angele, Stefan Decker, Michael Erdmann, Hans-Peter Schnurr, Steffen Staab, Rudi Studer, and Andreas Witt. On2broker: Semantic-Based Access to Information Sources at the WWW. In *Proceedings of the World Conference on the WWW and Internet (WebNet 99)*, Honolulu, Hawaii, USA, October, 25–30 1999.

[6] Dublin Core Metadata Initiative. homepage of the Dublin Core Metadata Initiative. *http://purl.oclc.org/dc/*, 1999.

[7] National Library of Medicine. homepage of the Unified Medical Language System (UMLS). *http://www.nlm.nih.gov/research/umls/*, 1999.