Marking the National Curriculum - a New Model for Semantic Mark-up

Fredrik PAULSSON[1], Jonas ENGMAN[2]
[1]*Royal Institute of Technology, Lindstedtsvägen 5, Stockholm, SE-10044, Sweden*
Tel: +46 (0)70 5605358 Email: frepa@nada.kth.se
[2]*CINS, Tvistevägen 47, Umeå, SE-90179, Sweden*
Email: jonas@cins.se

Abstract: The use of information structure and metadata for the National Curriculum becomes more and more crucial as the demands for advanced eLearning services increases. In a Swedish pilot project a simple information model was developed and tested on the National Curriculum. An implementation of a common XML based Mark-up Language called TEI was used for the first approach. An alternative model where three different kinds of metadata where identified and separated using RDF, XML and Annotea as core technologies was used for the second approach. Two approaches where tested and evaluated against five criteria. It was found that XML-based mark-up languages like TEI fail in cases where complex metadata and a combination of several metadata models are required. In conclusion we argue for the importance of separating information structure, descriptive metadata (including semantics) and catalogue metadata.

1. Introduction

Learning and teaching are going through extensive changes as the use of new technology is increasing rapidly. With this in mind it is important that national authorities are sensitive to the development of IT and learning. As the use of digital learning content increases, national authorities must be prepared to provide content of national relevance – not only in digital formats – but in digital formats usable for learning, so that it can be used to improve and enhance the work of teachers and students and at the same time be integrated into daily school activities. To date digital publishing is often limited to unstructured digital formats (e.g. html, .doc). This is not sufficient to meet future demands for information structure and metadata that is a condition for sophisticated e-services for eLearning.

In Sweden, as well as in many other countries, the national curriculum is among the most central documents provided by national authorities. But the national curriculum doesn't stand alone and the term "National Steering Documents" (NSD) [2] will be used to describe the documents that constitute the foundation for the Swedish school system. The national steering documents consist of: the Education Act, school curriculum, program objectives, course syllabi and grading criteria [2]. The Swedish school system is regulated by goals and guidelines set by the government, implemented in schools (preferably by the adoption of local steering documents) by the municipalities and evaluated by the National Agency for Education. The steering documents describe two kinds of goals for each subject area: goals that must be reached and goals to strive for. Those goals are the most central parts of the NSD. As a consequence the Swedish school system is decentralized in its organization by most aspects. The National Agency for Education expresses that: *"Part of the philosophy of the Swedish education system is that the state should define the national goals and guidelines for education, while detailed sub-regulation should be avoided to give municipalities and schools as much freedom as possible to formulate their own work."* [2]

For this reason it is essential that the national steering documents are made available and that the national steering documents can be used when local steering documents are formulated as well as in connection to classroom activities. To facilitate such use, the national steering documents need to be available in a digitally useful format and furnished with relevant metadata and markup. This need has been emphasized by the increasing focus on individualization and quality in learning. One of the most salient driving forces is the introduction of individual development plans (IUP) [11] in Swedish schools. The use of IUP is suggested to be prescribed by the upcoming school law. IUP is a method for dealing with individual knowledge goals as well as the origins of the knowledge goals and for visualizing progress and quality. This makes the connection between the IUP and the goals in the local- and national steering documents vitally important. The local steering documents are derived from the national steering documents and it is essential to point out the relations between them in order to put the local steering documents into context. Equally important is the need to make a connection between Steering Documents and learning content. It is often desirable to associate learning content (such as Learning Objects [26]) with a specific goal or wording in the NSD. Most modern learning content repositories use metadata to describe Learning Content. By adding information structure and metadata to the national steering documents it will be possible to filter and match different sets of metadata to establish- and visualize relations between Learning Content and NSD - and vice versa. Additionally it is desirable to establish relationships between Steering Documents for different subjects in order to facilitate thematic work and thematic use of learning content. [16]

1.1 Delimitations and Disposition

This paper describes a case - the Swedish National Curricula Markup pilot project (NCM) - where three of the Swedish NSD where provided with information structure and metadata markup. This paper describes the evaluation and development of methods for metadata with a focus on the problems and challenges encountered when working with digitalization and markup of national documents. The primary target groups for the NCM-projects were civil servants at the national school authorities and teachers. Secondary target groups are producers of digital learning content and local policy makers.

Next section describes some previous work related to metadata markup and curriculum. After that the methods and activities of the NCM-project is described, followed by the conclusions and proposals for further work.

2. Previous Work

Some efforts have been made in this area, such as the British Curriculum On-line project [3]. Curriculum On-line has a slightly different focus to the NCM project as it is focused on how to integrate IT effectively into the classroom by pointing out high quality learning resources and emphasizing the relationship between curricula and learning resources [3]. BECTA Vocabulary Management Tool is another UK initiative for managing vocabularies [6]. In Norway, Topic Maps is used to describe ontologies in the OLUF project.

A lot of work has been done for metadata and markup of digital resources. Both for descriptive and catalog metadata [6] [19] and content markup such as the Text Encoding Initiative [10] [5] as well as for structural metadata markup [1] and for the use of metadata in public governmental organizations [22]. Specific standards for educational metadata exist, usually based on the IEEE Learning Objects Metadata, LOM [14]. There are two main technologies for implementing semantic web metadata [4], the W3C Resource Description Framework RDF [9] and the ISO Topic Maps [18].

3. Objectives

The overarching problem faced by the NCM-project was to discover methods and technologies that enables digital distribution of NSD in a structured way and with metadata- and structural mark-up that allow the needed flexibility and at the same time ensures that the all the rules and regulations for public documents are met in a satisfying way. One of the main purposes is to allow for third party to use NSDs to provide new services for schools. It is therefore important that the resulting model supports several layers of metadata for a resource, as well as metadata according to several metadata models or Application Profiles [9]. Five criteria that must be fulfilled were identified for those purposes: (1) The content, semantics and meaning of the original document must be preserved; (2) The format for data distribution must be structured, open and application neutral. (3) Markup of a whole resource as well as of a portion of a resource must be supported. (4) Several metadata models and application profiles must be supported for the same digital resource (as well as parts of a resource). (5) Possibility to add markup without interfering with the original document or existing metadata.

4. Methodology, Technology and Implementation

The main activities during the NCM-project where: a survey of existing models and methods for structure- and metadata mark-up, development of a simple metadata test-model for the Swedish national steering documents, a prototype implementation of the metadata model and an evaluation of the prototype against the five criteria. As the evaluated prototype was found to be too limited and a second prototype was implemented using an alternative approach that better suited the five criteria.

4.1   The Metadata Model and TEI

The survey of existing models made it clear that the closest match among established markup language technologies was TEI [10]. It was decided that TEI should be used for a first prototype to test the metadata model and the concept of content markup of NSD. It was however clear from the beginning that TEI would not be the final and most suitable solution since it was clear from the start that it fails to conform to crucial criteria.

In a joint venture, involving civil servants and researcher a simple metadata model, suitable for TEI, was developed. The model included only the most basic elements needed for curricula markup. The metadata was organized into categories based on the most central concepts. Similar concepts where grouped into "boxes" containing "markup collections". An example of such category is "KREA" (for "Creativity"; containing: Activity, Creativity and Fantasy). Listing 1 show an example of an extract of a paragraph from the TEI/XML marked NSD for Swedish language. Line 9 of the example shows how the Swedish word "fantasi" (fantasy) is marked using the markup category "KREA", using the keyword FANTASI.001 (fantasy.001).

Listing 1 shows parts of the structural markup. Some of it, such as <p> (paragraph), <list> (list) and <item> (item) are given by the TEI DTD, others such as <div atype="goal-strive"> in line 1 are defined by the NCM model. Such markup can be regarded both as structure and as descriptive metadata carrying semantics. It describes a subdivision in the structure, but tells us at the same time that this subdivision is a "goal to strive towards" which has a clearly defined semantics defined by the National Agency for Education, but in another context the semantic may be slightly different. The first prototype was implemented using existing Open Source software. The Apache Xindice Database was used to store the TEI/XML files. Apache Xindice is optimized for storing and searching XML. Apache Xalan was used for XSL transformation, together with Formatting Objects Processor (FOP)

and Cascading Style sheets for presentation. The functionality was fairly limited: the national steering documents could be viewed as html or PDF, possibility to search for a concept or a theme in the NSD or for a specific subject or theme in the NSDs for two or more subjects. The resulting output could either be the NSD(s) with the matching concepts highlighted or an electronic compendium, which is a compilation of extracts dealing with the concepts in question.

```
1  <div atype="goal-strive">.
2  <head>Mål att sträva mot</head>.
3      <p>Skolan skall i sin undervisning i <rs atype="concept" key="SVENSKA.016"
       reg="KULT">svenska</rs>.
4      sträva efter att <rs atype="concept" key="ELEV.088"
       reg="LARPROC">elev</rs>en </p>.
5      <p>.
6      <list>.
7          <item>.
8              <rs atype="concept" key="UTVECKLA.033" reg="KREA">utveckla</rs>r sin
9              <rs atype="concept" key="FANTASI.001" reg="KREA">fantasi</rs> .
10             och lust att lära genom att <rs atype="concept" key="LASA.002"
               reg="SPRAK">läsa</rs>.
11             <rs atype="concept" key="LITTERATUR.008" reg="ARTS">litteratur</rs>
12             samt gärna läser på egen hand och av eget intresse,.
13         </item>.
14     .
15         <item>.
16             <rs atype="concept" key="UTVECKLA.034" reg="KREA">utveckla</rs>r sin
17             <rs atype="concept" key="FANTASI.002" reg="KREA">fantasi</rs> och
               lust .
18             att skapa med hjälp av <rs atype="concept" key="SPRAKE.005"
               reg="SPRAK">språke</rs>.
19             t, både individuellt och i samarbete med andra, .
20         </item>.
```

Listing 1 A TEI marked portion from the NSD for Swedish language.

The evaluation of the first prototype focused the ability fulfill the five criteria. It showed that TEI met the second and third criteria. TEI partly fulfill the forth criteria since it is theoretically possible to compile two different metadata models in the TEI/XML-markup, but as seen in listing 1 the XML tends to get quite messy and it actually becomes more markup than content. More serious problems occur in cases where overlapping markup is needed. Overlapping markup is must be used when more than one set of metadata is needed for describing the same portion of a resource and it will no longer be possible to have well formed XML with overlapping tags. The first prototype suffers from serious problems when handling more that one model for metadata describing the same resource. The prototype did not at all meet the second and fifth criteria. The main reason for this shortcoming is that TEI expresses metadata semantics using XML markup inside of the document. Since TEI puts its metadata inside the same XML file that contains the document it becomes impossible to add metadata without altering the original file.

## 4.2   Second Prototype: the Annotea Model

The first prototype pointed out the importance of separating markup describing information structure, metadata and semantics for content markup (describing a portion of the resource) and metadata for cataloging resources. For this reason the second prototype was developed using an alternative model for managing metadata. The new model makes a clear distinction between three different kinds of metadata: structural markup (XML), metadata that refers to a portion of a resource and catalog metadata describing the whole resource. The implication of the new model was that all metadata, except structural metadata, was separated from the resource. As a consequence, the resource was not altered besides the addition of basic structural XML-markup - which can be regarded as a part of the original semantics as it describes the structural elements of the original resource.

The strategy for the second prototype was to use established, open technologies stitching them together using as little system development as possible. RDF was chosen for

the metadata implementation. The choice of RDF was mainly motivated by the needs described in [17] and [16] (as referred in the introduction to this paper) as well as by criteria two and four. One of the main problems when using RDF concerns efficient storage of RDF tipples in order to take advantage of the full potential of RDF. To solve this problem the Standardized Contextualized Access to Metadata (SCAM) was used for storing and accessing RDF-metadata. Issues concerning storage, management and access of RDF metadata are discussed in detail in [17] and [15]. The "Standardized Hyper Adaptable Metadata Editor" (SHAME) [15] was used for adding functionality for metadata editing. SHAME is a developer's framework for metadata editors, metadata presentations and query interfaces for RDF metadata. The Simplified DocBook DTD was used as the format for storage and structure markup of NSD [24]. Simplified DocBook was chosen in order to fulfill criteria two. Simplified DocBook descends from DocBook [25], which is a well-established markup language for text documents. It would have been possible to continue to use TEI, but since there was no need for TEIs "semantic" capabilities the choice fell on simpler markup language. DocBook is suitable for XSL-transformation [20] and a lot of ready-made style sheets are available. In theory, almost any XML based markup language could be used together with the new model for metadata management and the choice of DocBook may very well be in reconsidered in the future.

The rest of the technical settings consisted of several of Open Source software: Apache Xindice was used as a native XML-database, based on its capabilities to store and search in XML files as well as its capabilities to do XPath and XPointer queries. The XLiP framework (from Fujitsu) was used to address XPointers [21] into XML documents. The XLiP is a fully compliant XPointer implementation, built upon the standard DOM API, which makes it a perfect match with the XML database. A JBoss application server and an Apache TomCat were used to run the different application components in the system.

The World Wide Web Consortium (W3C) Annotation technology Annotea [23] was one of the most central technologies in the second prototype. Annotea considers all annotations to be RDF metadata [12]. The Annotea protocol can be used for annotations, using existing Annotea schemas, as well as for general-purpose metadata [13]. An Annotea annotation is a RDF statement that's "hocked-on" using standard XML technologies such as an XPointer to point into the part of the document that is of interest. RDF is used to identify the resources using an URI (in this case an XPointer) and to make a statement about the resources (regarded as the subject), using properties and values. In this case the resource is a NSD (or a part of a NSD) at which the URI points. The subject must be a resource, i.e. something we can identify with an URI. All properties must also have an URI.
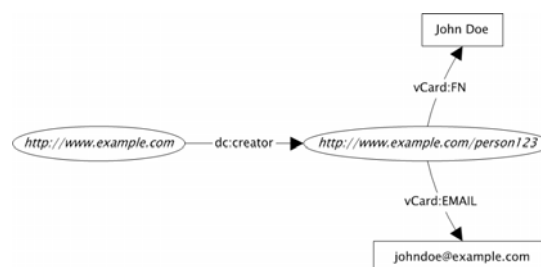
Figure 1 John Doe is represented as a resource. A vCard is used to specify his name and his e-mail.

If the property has an URI it becomes certain that the property is unique. If two properties that has the same URI they are in fact the same property and can thus be processed in the same way. A RDF object can be another resource or a constant value. This means that a resource can be described using a manifold set of metadata where several values are expressed (figure 1). This becomes even more complex by using RDF Containers to

describe grouped properties with rich semantics [8]. As Annotea addresses metadata by XPointers, the metadata becomes separated and independent of the resource. Annotations can be regarded as RDF/XML that "points" into the Annotated resource. When using the Annotea Annotation Schema [23] to construct annotations, every annotation has its own XPointer (as a context property). The pointer can either point to a single XPointer point such as an XPath node-subtree, or a range which may start and stop on arbitrary points in the XML document. Figure 2 shows an overview of the system. The server acts as an Annotea client and fetches the document from the XML store. The Annotea client issues a request for all annotations that annotates the document. The request is a standard Annotea compliant request [23], which is translated into a SCAM metadata repository query. The resulting RDF is returned to the Annotea client and the annotations are merged into the XML document. The XML document is inserted into an XHTML skeleton that enables the use of scripts on the resulting page.
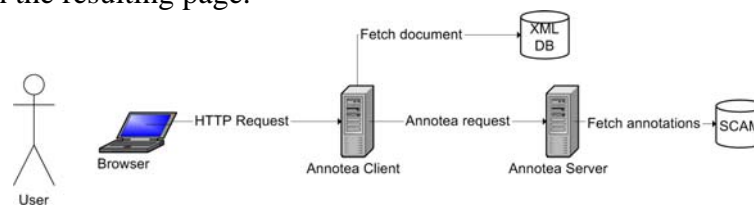

Figure 2 System Overview.

When a user adds or updates an annotation, an ECMA-script in the browser issues a HTTP POST to the Annotea client containing an XPointer that points at the user specified location. The Annotea client responds with a HTML form that is generated by SHAME. The form creates or modifies the underlying RDF structure of the annotation. When the form is posted to the Annotea client, an add-operation is forwarded to the Annotea server. The server then adds or updates the RDF graph into SCAM.

5. Results

The second prototype fulfills all five criteria by supporting a model where the three identified types of metadata is clearly separated (figure 3). The two types of descriptive metadata are stored in different metadata stores and "pointed in". The model allows for any number of actors to provide metadata independent of each other and without interfering with the resource or existing metadata. This separation enables a flexible use of metadata Different types and sets of metadata from different providers can be combined, filtered and used in various ways.
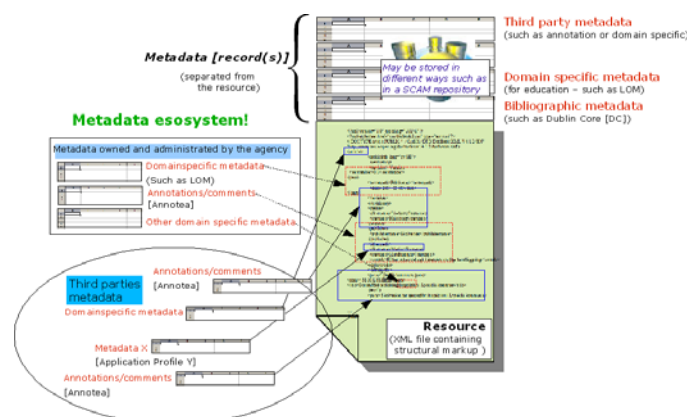

Figure 3

The semantic capabilities of RDF can be exploited from primitive metadata use to advanced machine reasoning – depending on the ambition and context. This kind of flexibility will

become more important as the use of metadata evolves and becomes more sophisticated - a reasonable scenario for the future. The project shows that Semantic Web Technology can be used to solve complex metadata issues.

6. Conclusions

In conclusion we argue that it is necessary to separate the different layers of the information structure to obtain the flexibility and functionality required by the five criteria, obtaining a sustainable system for advanced metadata management. It is of great importance to separate structural mark-up from descriptive and semantic mark-up to facilitate metadata layering. Then metadata can be provided by different actors, transparently independent managed and using different metadata models. Mark-up languages like TEI have a niche in their specific domains, but are not suitable for managing advanced general-purpose metadata. We believe that the Annotea-based approach to metadata mark-up is well worth exploring further and that it is applicable to other domains and subject areas as well.

An interesting possibility is the ability to combine NSD markup with metadata from other sources - such as repositories for digital learning resources. A small test was made as where the NSDs where connected to a repository called NoT-navet (a repository for Science resources) with good results. This gave the possibility to filter learning resources metadata using parameters that was dynamically extracted from NSD metadata. It was accomplished by combining the original search parameters from the NSD with synonyms and semantically related concepts resulted from the filtering of the NSD metadata. Those where then used to generate a search string that was sent to NoT-navet. Suddenly it became possible to find learning resources that were originally intended for science but which where of interest for history teaching for instance. Further work is needed in this area, since the small test made in the NCM project was fairly unsophisticated.

Further work is also needed to take the system from the prototype to production stable. It is estimated that 12-18 person month of additional system development is needed to take the prototype to production stable. When production stable the code will be distributed as Open Source.

Most of the used technologies are stable, but there are still some uncertainties on how (and when) to use the different technologies. It is also unclear what happens if an Annotea marked XML file is changed. Some simple tests shows that minor changes are handled, but that major changes may cause problems. This must be encountered for.

References

[1] Abiteboul, S., Buneman, P., & Suciu, D. (2000). Data on the Web : from relations to semistructured data and XML. San Francisco, Calif.: Morgan Kaufmann.

[2] About the national Agency for Education. (2004, 2004-01-28). Retrieved 2004-12-07, 2004, from http://www.skolverket.se/english/about.shtml

[3] BECTA. (2005). Supporting learning and teaching in secondary schools - Curriculum Online. Coventry: British Educational Communications and Technology Agency.

[4] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American.

[5] Burnard, L., & Sperberg-McQueen, C. M. Text Encoding Initiative (Projekt), (1994). Guidelines for electronic text encoding and interchange (pp. xxvi, (1290)). Chicago: TEI.

[6] Collett, M. (2005). Becta Vocabulary Management Tool. Retrieved 1 April, 2005, from http://www.estandard.no/docs/voc_tools/Vocab_manager_brief_v0_3.pdf

[7] Haynes, D. (2004). Metadata for information management and retrieval. London: Facet.

[8] Hayes, P. (2002). RDF Model Theory. Retrieved February 14, 2003, from http://www.w3c.org/TR/rdf-mt/

[9] Heery R, P. M. (2000). Application profiles: mixing and matching metadata schema's - introduce the 'application profile' as a type of metadata schema. Ariadne (25).

[10] Ide, N., & Véronis, J. (1995). Text encoding initiative : background and context. Dordrecht: Kluwer Academic.

[11] Individuell planering och dokumentation i grundskolan. (Report No. Dnr 2003:251)(2004). Stockholm: National Agency for School Improvement.

[12] Kahan, J., Koivunen, M.-R., Prud'Hommeaux, E., & Swick, R. R. (2001, 1-5 May 2001). Annotea: An Open RDF Infrastructure for Shared Web Annotations. Paper presented at the WWW10 International Conference, Hong Kong.

[13] Koivunen, M.-R., & Swick, R. (2001, October 21, 2001). Metadata Based Annotation Infrastructure offers Flexibility and Extensibility for Collaborative Applications and Beyond. Paper presented at the KCAP 2001 workshop on knowledge markup & semantic annotation, Victoria B.C., Canada.

[14] LOM Draft Standard. Final 1484.12.1-2002 (2002). IEEE: IEEE-Standards Association.

[15] Palmér, M., Naeve, A., & Paulsson, F. (2004). The SCAM Framework: Helping Semantic Web Applications to Store and Access Metadata. Paper presented at the European Semantic Web Symposium 2004, Heraclion Greece.

[16] Paulsson, F. (2004). Märkning av nationella styrdokument (pp. 8). Stockholm: National Agency for School Improvement.

[17] Paulsson, F. (2003). Standardized Content Archive Management – SCAM. IEEE Learning Technology newsletter, 5(1), 40-42.

[18] Pepper, S., Vitali, F., Garshol, L. M., Gessa, N., & Presutti, V. (2005, 29 March, 2005). A Survey of RDF/Topic Maps Interoperability Proposals. Retrieved 2 April, 2005, from http://www.w3.org/TR/rdftm-survey/

[19] Powel, A. (2003). Expressing Dublin Core in HTML/XHTML meta and link elements. Retrieved December 12, 2004, from http://dublincore.org/documents/dcq-html/

[20] Stayton, B. (2005). DocBook XSL: The Complete Guide (3rd ed.): Sagehill Enterprises.

[21] Simpson, J. E. (2002). XPath and XPointer : locating content in XML documents (1. ed.). Beijing ; Cambridge ; Farnham ; Köln ; Paris ; Sebastopol ; Taipei ; Tokyo: O'Reilly.

[22] Song, W. W. (1999). Metadata for the management of electronic documents in the governmental organizations and learning objects in the learning domain (Research report No. SITI 99:03). Kista: SITI, SISU.

[23] Swick, R., Prud'hommeaux, E., Koivunen, M.-R., & Kahan, J. (2002). Annotea Protocols, from http://www.w3.org/2001/Annotea/User/Protocol.html

[24] Walsh, N., (ed.). (2004). The Simplified DocBook Document Type, Working Draft 1.1CR1, : The Organization for the Advancement of Structured Information Standards [OASIS].

[25] Walsh, N., & Muellner, L. (1999). DocBook : the definitive guide. Beijing ; Farnham: O'Reilly.

[26] Wiley, D. A. (2002). The Instructional Use of Learning Objects. In D. A. Wiley (Ed.). Bloominton: Agency for Instructional Technology and Association for Educational Communications & Technology.