Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 257-259

A Proposal for a Flexible Validation Method for Exchange of Metadata between Heterogeneous Systems by Using the Concept of MicroSchema

Jens Vindvad
Riksbibliotektjenesten, Oslo Norway
Jens.Vindvad@rbt.no

Erlend Øverby
Conduct AS, Oslo Norway
erlend.overby@conduct.no

1. Introduction

A new method to solve the validation problem that arises when exchanging information between heterogeneous systems is proposed. The problem of validation is addressed by introducing the concepts of MicroSchema, used in a namespace environment.

To be able to share information between different systems, a well-defined protocol for information exchange must be in place. XML (Bray et al. 2000) has emerged as a new protocol for use in information systems for exchanging information between different systems.

Some of the challenges, when importing metadata from one system to another, are described in the experience learned by iLumina (McClelland et al. 2002) when importing IMS metadata. An issue reported was the need of validating against XML-model and error checking of imported metadata.

Normally two alternatives exist to describe and define the information structure or model in an XML document, the first is a DTD (ISO 8879) and the second is an XML-schema (Thompson et al 2001). Both these approaches currently have the disadvantages that in order to validate and check the structure of the information, description of the whole structure and all its possibilities and constraints must be in existence in one large and inflexible model, making it harder to establish an efficient validation of data exchange between different systems.

One reason almost everyone is using XML in only well-formed manner – is the flexibility in generating the information structures, if a new element is needed – it is just added and the information structure is still well-formed. Validation is often sacrificed. The disadvantage of only well-formed structures is that almost any element can be included, and there is no control of what the element names are or of their semantic meaning.

2. Conflict between rigid structures and the need for flexibility

When working with structured information, there is a conflict between flexibility, and the need for a rigid structure. If we try to look at the structure we normally find in a book, we will se that in many of our content models there is many similar structures. Normally parameter entities is used to manage that flexibility, but there is still a need to change the structure and to create new version of the DTD's. When using schemas to describe the structures, the notion of "global" element definitions can be used, but there is no function for describing content models in a flexible and reusable way. If wanting to change a content model by adding some new elements, is has to be done in a many different places in a schema, and only at once in the DTD.

One of the nice new features with XML over SGML is the introduction of the Well Formed document – which has the implications that there is no need to have a specified structure defined for the XML-document. This gives a great flexibility in processing the XML-documents and normally this is sufficient when there is full control of the information, and the processing of it. But if several people or systems producing information there is a need for greater control over the structure of the information that is produced.

3. MicroSchema

The challenge is to combine the flexibility in the well-formed document, with the control of the valid document. Using MicroSchema's this flexibility can be provided. The idea of a MicroSchema is that it should only describe a very small piece of information, and only such information as is relevant to the

specific description. Information that is not relevant to the specific context is described in another schema. MicroSchemas combines the flexibility of only well-formed documents with the need to specify and validate complex structures. To be able to express the relevance and the connection between MicroSchemas, a standard method of enhancing the schema specification in order to address the valid elements in the specific context is needed. Using namespaces, introducing the term "Allow-schema-namespaces", will do this.

Instead of specifying the whole structure in one or more schemas, only a small part of the structure in its own Schema (MicroSchema) is specified. Then the URI's is used to specify parts of the flexible Content Models. To some extent Parameter Entities can be looked upon as a URI reference from the MicroSchema. And the specification of Content Model or of the Generic Identifier (GI) is defined at the target URI. The URI will also work as the Namespace specification of the Semantic meaning of the GI's.

MicroSchema URI can bee addressed in two ways; one is as the Content Model specification, where one specific MicroSchema file is addressed in the URI.

```
<xs:element name="*" msc:gi="http://www.rbt.no/xmlns/
cerif/output/misc/chapter.msc"/>
```

Example 1 Using the MicroSchema attribute GI

In example 1 the xs:element will get the GI and Content Model of the element specified in the MircroSchema addressed at the URI http://www.rbt.no/xmlns/cerif/output/misc/chapter.msc. At the other hand only the Content Model could also be specified, using the MicroSchema specification for one element as shown in example 2.

```
<xs:element name="kapittel" msc:cm="http://www.rbt.no/
xmlns/cerif/output/misc/chapter.msc"/>
```

Example 2 MicroSchema specification for one element

In example 2 the element name "kapittel" will get the same Content Model as the MicroSchema specified at the given URI. Here this will replace the CHAPTER GI specified in the chapter.msc MicroSchema with the GI KAPITTEL given as the value of the name attribute.

The MicroSchemas and the corresponding documents are valid XML documents, and therefore can be processed as such. One of the primary ideas behind the MicroSchema is the XML-Well-formed processing, which does not require a set of rules against which to check the structure of the information. All XML MicroSchema documents are at least well-formed. The idea of a MicroSchema is to have the possibility of combining both well-form-ness and strict structures where the structure is expressed in a MicroSchema. Introducing the following three forms of MicroSchema processing rules does this: simplest form, simple MicroSchema check and complete MicroSchema validation.

4. CRIS as a test case

A lot of work has been done in the field of metadata exchange. Particularly initiatives like Dublin Core, Open Archive Initiative and work with Learning Object Metadata (LOM). To demonstrate and test the concept of MicroSchema a new flexible XML-model for exchange of research documentation in Current Research Information Systems (CRIS) has been developed and proposed. A working XML-exchange model for metadata exchange between different CRIS and between with library systems and CRIS have been tested. A technical report describing the test case will be published summer 2002, the title of the report is: "Technical report of June 2002. Proposal for a flexible and extensible XML-model for exchange of research information by use of MicroSchema : Description of a working model for documentation produced by researchers".

5. Conclusion

A more flexible approach is needed to validate the exchange of data between different information systems. To solve this need, the concept of MicroSchema is introduced.

A new flexible and extensible XML-model for exchange of research information is proposed, using MicroSchema. The new XML-model has been tested against existing CRIS-systems, and data has been successfully imported into the model. The model has also with success been tested against ordinary library catalogue data.

References

Biron, P.V. and Malhotra A., eds. 2001. XML *Schema Part 2: Datatypes*. The World Wide Consortium (W3C) http://www.w3c.org/TR/2001/REC-xmlschema-2-20010502

Bray, T.; Paoli, J.; Sperberg-McQueen, C.M. and Maler, E., eds. 2000. *Extensible Markup Language (XML) 1.0 (Second Edition)*. The World Wide Web Consortium (W3C). http://www.w3c.org/TR/2000/REC-xml-20001006

Fallside, D.C., eds. 2001. XML Schema Part 0:Primer. The World Wide Consortium (W3C) http://www.w3c.org/TR/2001/REC-xmlschema-0-20010502

ISO 8879:1986 *Information processing Text and office systems – Standard Generalized Markup Language (SGML)*

McClelland, M.; McArthur, D.; Giersch, S. and Geisler, G., 2002. Challengs for Service Providers When Importing metadata in Digital Libraries. In: *D-Lib Magazine*, 8 (4)

Thompson, H.S.; Beech D.; Maloney, M. and Mendelsohn, N., eds. 2001. *XML Schema Part 1: Structures*. The World Wide Consortium (W3C) http://www.w3c.org/TR/2001/REC-xmlschema-1-20010502