Proposals for a new flexible and extensible XML-model for exchange of research information

Jens Vindvad, Erlend Øverby
National Office for Research Documentation, Academic and Special Libraries, Oslo
Conduct AS, Oslo

Summary

In this paper a new flexible extensible XML-model for the exchange of research documentation is proposed, and a working XML-exchange model is described. The working model is limited to documentation produced by researchers. The ideas, XML-model and construction proposed and used in the working model are extensible, and can be expanded to the whole field of research documentation, and to other fields as well.

The paper is partly based on a report to be completed in May 2002. That report will provide a full description of the model, which is not possible in a short paper presented at the CRIS2002 Conference. The report aims to set out the groundwork and facts, to document the proposed new XML-exchange model, and to contribute to the further discussion in euroCRIS in respect of data exchange between different systems.

1 Introduction

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 used in information systems for exchanging information between different systems.

This introduces a new problem of how to specify the information structure in XML. Our proposal addresses this problem. When exchanging information between different systems, the system needs to know what type of information it receives, and the structure of this information, so that the system is able to transform the information into the specified system. Normally when information is exchanged XML, a DTD (ISO 8879:1986) or an XML-Schema (Biron & Paoli 2001) is specified to describing the structure of the exchanged information. However, creating an XML document from existing information that is valid with the XML-schema could be as difficult as transforming your information into a specified protocol.

One example of this situation is taken from Scotland (Woolman 2001) where a large-scale program is under way, designed to develop an XML schema for electronic clinical communications. Another example is reported from the chemical community (Murray-Rust et al. 2001) of an operational system for managing complex chemical content entirely by means of an XML-based markup language called "Chemical Markup Language (CML)" where an XML schema has been developed.

Two alternatives exist to describing the defined structure of information; the first is a DTD and the second is an XML-schema. Both these approaches currently have the disadvantages that in order to validate and check the structure of the information, a description of the whole structure and all its possibilities and constraints must be in existence. This makes the exchange model large and inflexible, which makes it harder to establish an efficient exchange of information between different systems.

Normally validation is sacrificed for well-formed structures. The disadvantage of well-formed structures is that they could include almost any element, and there is no control of what the ele-

ment names are and what their semantic meaning is. In our proposal we try to address and solve this problem by introducing the concept of *micro-schemas*.

2 Fundamental concepts

2.1 XML-exchange model

In many of the structured systems based on SGML (ISO 8879:1986) - and to some extent XML - the structure of the information is vital where the structure of the information is described in one DTD. And for the documents to be valid with the DTD - all possible information within the document/information has to be described in that DTD. This means all the information in all contexts has to be described in that context in the DTD. As a consequence of this complexity the DTDs are hard to understand and to use.

To solve this problem we have introduced a new way of looking at complex information structures - described in XML - using a much more flexible approach to the possible ways of describing information structures, and to ease the exchange of information between information systems.

In a recently published paper (McClelland et al. 2002), the exchange of metadata between digital libraries is discussed. The final conclusion is: *"This article outlines some of those challenges and underscores the point that not all the problems will be solved merely by adopting a common metadata element schema."*

This would seem to support the view that there is a need for a more flexible approach than which is currently used, e.g. by using a micro-schema.

2.2 Architectures

Since the information domain is known - Current Research Information Systems (CRIS) - the nature of the information is somewhat similar. Different systems describe the information in different ways - but since the information is in the same information domain, the information model is to some extent the same and is often based on CERIF2000 (Alexandraki et al. 1999).

The architecture forms the building blocks to describe the information models, used when specifying the exchange model.

To ease the exchange of research information, the information should be compliant to an exchange architecture that is flexible, but where all the information elements are well defined, and with a vocabulary agreed upon.

To be able to exchange information, the existing data model needs to be transformed into the exchange architecture model. Similarly, to receive information, the exchange architecture model must be transformed into one's own data model.

By using this idea of an architecture specified in micro-schemas, it will be easy to adopt and modify the model for other types of information, and perhaps this model could ease the exchange of information between other systems as well. The micro-schemas will be the building blocks of the information exchange structure.

2.3 Vocabularies

To make the exchange model work it is mandatory that the vocabulary be agreed upon, in order to achieve a unique and well-defined semantic meaning of the information element. This view is supported in a recently published article about metadata (Duval et al. 2002) where one of the conclusions is: *"For these opportunities to be realized, some convergence of encoding formats and commonly agreed semantics will be necessary."*

2.4 Micro-schema

Our idea is to address the specification of the smaller information elements, in order that the logic information can be transformed to these smaller information elements, with the system able to check the validity of these smaller information elements.

The idea is that the export and import system knows the structure of these information elements, and can then easily encode the information using these smaller specifications, without taking into account the overall information model. The information elements are then mapped to the specified data structure. Since the information is within the same information domain, the natural structure of the information should be easily identifiable, even if the overall information model is different.

The import system will then validate the smaller information models against the micro-schema, and transform the information into the structure of the import system. This ensures a flexible and open method of exchanging information between systems using XML as the information medium.

This proposal also specifies how to parse and validate information using these micro-schemas, and how to enhance the model by specifying new micro-schemas.

2.4.1 Addressing of the micro-schema

To be able to address the micro-schema used, namespaces are employed. At the given namespace URI, it is expected that a schema will be found, which will be parsed by the export/import system. If there is no such schema, the information will only be well formed, and it cannot be ensured that the exchange information conforms to the exchange model. Each micro-schema may address other micro-schemas by means of a new proposed syntax in the schemas by using their namespace URI. The system must then look up this schema and parse it to check the validity of these structures.

3 Step-by-step description of how to achieve the new XML-exchange model

To achieve the new XML data exchange model the following eight steps are proposed:

3.1 Establish an extensible framework

The working model is limited to information produced by researchers. Information produced by researchers can be described in the following framework:

3.1.1 Outputs

The collection of all types of information produced by the researchers is called output. Outputs are divided into four subgroups: - results, - communication, - documentation and - art.

3.1.2 Results

Results are taken to mean the results of the research produced by the researcher in person. Examples of results are: publications, patents and products.

3.1.3 Communication

By communication we want to label the forms of communication that researchers use in their work. Researchers often need to or wish to discuss their ideas and views. This form of communication is not a result of their work, but represents interesting and important steps in the process of producing results. Such communication has two main audiences: the professional community

and the general public. Examples of communication are: conference presentations, workshops, broadcasting and interviews in the press.

3.1.4 Documentation

A researcher has to carry out administrative tasks and produce documentation, which cannot be classified as results or forms of communication. This can be pure administration or high level of professional work. In connection with CRIS systems documentation is not always the most interesting part, but it is a necessary element to give a complete picture of the information produced by a researcher. Examples are: reports to funding institution, application for funding, documentation of a laboratory upset, administrative tasks of a research projects, and computer programs.

3.1.5 Art

Art is not a necessary output of a researcher's work but it can be. Art can be seen as a result in itself, a form of communication or type of documentation or all of these. Art needs and deserves a classification based on standards used and accepted in the art community. Examples of art are: works of art, exhibition and performance.

3.2 Establish an internal information structure

It is necessary to establish an internal information structure. The internal information structure will be used to define micro-schemas and to build information containers. This step also has to be seen in relation to the next step, wording and vocabulary, and the first step of framework. The internal information structure is the basis the new XML-model will rest on.

3.3 Propose wording and establish a vocabulary

Successful communication requires common wording and definitions. For an exchange data model to be established, a definition has to be agreed upon. In the report, suggestions for definitions of all used terms are given. A lot of the terms employed are commonly used words. Precise definitions are required, and for some terms a taxonomical approach is used, since this permits testing to see if a concept or term belongs to a particular definition.

Space does not permit definition of all terms given in this paper. The following is an example of how it is proposed to provide a definition of the term "publication".

3.3.1 Publications

Publication is a commonly used word, which in daily use does not have a precise definition. To establish a vocabulary and a namespace, we need the word "publication" and have to give it a precise and distinct definition. To do this we establish the following five tests, which must be complied with before we call an information unit a publication. The five tests involve: - *addressee*, - *copies*, - *location*, - *readability* and - *time*.

Addressee: For a work to be a publication the general public has to be the addressee of the publication. A publication cannot have an explicit addressee. If the publication is addressed to a limited group or single identity it is not a publication.

Copies: For a work to be a publication it must be available in several equal copies. It should not be possible to change history by changing or destroying the source of a publication or a few copies.

Location: For a work to be a publication it should in principle be possible to acquire a copy of the publication all over the world from any library that has access to the international library community network.

Readability: A publication should be readable without the use of technical equipment. Either the original publication must be readable without the use of technical equipment or a copy must be obtainable by transforming a publication into a format that permits reading without the use of technical equipment.

Time: For a work to be a publication the user should have the option to access and read the publication at any time the user wishes.

3.3.2 Test for addressee, copies, location, readability and time

The five tests which involve: - *addressee*, - *copies,* - *location*, - *readability* and – *time* are also used against the terms communication and documentation, but with other requirements to pass the test.

3.4 Propose a namespace

Based on the two previous steps, the internal information structure and wordings and vocabulary, a namespace is proposed. Establishing a namespace is an important step in establishing an exchange model. This allows others to make programs and style sheets using the data exchange model.

3.5 Define reusable microschemas using the internal information structure

The idea of a micro-schema 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.

To be able to express the relevance and the connection between the micro-schemas we need to develop a standard method of enhancing the schema specification in order to address the valid elements in the specific context. Using namespaces, introducing the term „Allow-schema-namespaces“, will do this. The system then scans for possible new schema structures allowed at the specific point.

Each information model in the exchange model will be defined in its own schema definition. The system needs four different types of schema definitions: templates, context, phrases and inline.

3.6 Build information containers, which fits into the framework

Based on the defined micro-schemas, information containers are built. Such information containers must conform to the internal information structure and fit into the framework.

3.7 Testing

It is important that the concepts and a specific model are tested. Establishing test beds will make it easier for one community to exchange information with other communities. The creating of a working model should make it possible to prove the concepts. By testing the exchange between specific information systems, a check can be made to ensure that all information elements and the information structure are taken care of.

So far we have successfully imported real data form one of the main CRIS systems in Norway into the XML-exchange model. The work has demonstrated that properly structured information can easily be exported into the XML-exchange model. Tests have also been carried out using data from a library system. Our results have shown that certain minor information details are difficult to extract from a library system due to the cataloguing rules.

To carry out these tests, the data was transformed to the XML-exchange format with the aid of XSL (Adler et al. 2001) style sheets processed by an XSLT (Clark 1999) processor.

3.8 Extension rules

So far the model is not easily extensible. To make the model extensible the following rules are established:

- Rules for defining new micro-schemas.
- Rules for defining new information containers.

4 Operation of the XML-exchange model

When first using the exchange model it is necessary to set up a mapping from each system's information model onto the exchange system's architecture. The vocabulary specified in the exchange model is intended to give guidance in this mapping.

The mapping is best described using XSLT, which transforms your XML-data model into the XML-exchange data model.

At the conference a live demonstration of this model will be given.

5 Final section

5.1 Conclusion and advise for further work

A more flexible approach is needed to the exchange of data between different data systems. To solve this need, the concept of micro-schema is introduced.

A new flexible and extensible XML-model for exchange of research information is proposed, using micro-schema. The new XML-model has been tested against existing CRIS-systems, and data has been successfully imported into the model.

The model has also been tested with success against ordinary library catalogue data.

The new model needs more testing against other systems. Furthermore, the model requires further examination and discussion. If the ideas and direction outlined in the proposal are accepted, a way forward for further development and change has to be agreed upon.

6 References

Adler, S.; Berglund, A.; Caruso, J.; Deach, S.; Graham, T.; Grosso, P.; Gutentag, E.; Milowski, A.; Parnell, S.; Richman, J.; Zilles, S. (2001): *Extensible Stylesheet Language (XSL) Version 1.0.* W3C. http://www.w3c.org/TR/xsl

Alexandraki, M. et al. (1999): *CERIF 2000 Guidelines Final Report of the CERIF Revision Working Group.* DG XIII-D.4, European Commission. http://www.cordis.lu/cerif

Biron, P.V.; Malhotra, A. (Eds.) (2001): *XML Schema Part 2: Datatypes.* W3C. http://www.w3c.org/TR/xmlschema-2

Bray, T.; Paoli, J.; Sperberg-McQueen, C.M.; Maler, J. (Eds.) (2000): *Extensible Markup Language (XML) 1.0.* 2nd edition. W3C. http://www.w3c.org/TR/2000/REC-xml-20001006

Clark, J. (Ed.) (1999): *XSL Transformations (XSLT) Ver. 1.0.* W3C. http://www.w3c.org/TR/xslt

Duval, E.; Hodgins, W.; Sutton, S; Weibel, S. (2002): Metadata Principles and Practicalities. In: *D-lib magazine,* Vol. 8, No. 4

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

McClelland, M.; McArthur, D.; Giersch, S.; Geisler, G. (2002): Challenges for Service Providers When Importing metadata in Digital Libraries. In: *D-Lib Magazine*, Vol. 8, No. 4

Murray-Rust, P.; Rzepa, H.S.; Wright, M. (2001): Development of chemical markup language (CML) as a system for handling complex chemical content. In: *New journal of chemistry*, Vol. 25, No. 4, p. 618-634
Woolman, P.S. (2001): XML for electronic clinical communications in Scotland. In: *International journal of medical informatics,* Vol. 64, No. 2-3, p. 379-383.

7 Contact information

Jens Vindvad
Riksbibliotektjenesten
Tel. +47 23 11 89 00
Fax. + 47 23 11 89 01
e-mail: jens.vinvad@rbt.no
www.rbt.no

Postal Address:
P.O.B 8046 Dep
N-0030 OSLO
Norway

Visiting Address:
Kronprinsensgt. 9
Oslo, Norway

Erlend Øverby
Conduct AS
Tel. + 47 90 12 96 42
Fax. + 47 22 33 60 24
e-mail: erlend.overby@conduct.no
www.conduct.no

Postal Address:

P.O.B. 805 Sentrum
N-0104 OSLO
Norway

Visiting Address:

Biskop Gunnerus gate 2
Oslo, Norway