Semantic Documents

Example for CCTS-based and RDFa-annotated XHTML documents

Heli Lintula, Paula Leinonen, Virpi Hotti
School of Computing
University of Eastern Finland
Kuopio, Finland
heli.lintula@uef.fi, paula.leinonen@uef.fi, virpi.hotti@uef.fi

***Abstract*—If heterogeneous information systems store documents to the ontology-based document repository, they have to provide ontology-based documents (i.e. semantic documents). In this paper, we present how ontology-based documents (XHTML+RDFa) can be provided. The RDFa markup is used for embedding exchangeable information in the RDF format within the XHTML documents. Before we can provide ontology-based documents, we have to specify the data model. In this paper, we introduce how we modeled CCTS-based data and how we formed the RDF repository which supports the CCTS-based data. Our example is from the National Project for IT in Social Services (Tikesos) which main aim was to specify the ontology-based national archive to improve the interoperability of client data in diverse information systems.**

Keywords—data model, semantic document, Core Component Technical Specification (CCTS), Resource Description Framework (RDF), eXtensible HyperText Markup Language (XHTML), Resource Description Framework – in – attributes (RDFa)

I. INTRODUCTION

Large document repositories (e.g. archives) can benefit from ontologies which facilitate search and retrieval [1]. If several heterogeneous information systems store documents to the ontology-based repository, they have to provide ontology-based documents. In the ontology-based documents (i.e. semantic documents) the ontology is used for document annotation.

Work which promotes developing compatible social welfare information systems has been done in Finland since the 1990's. For example, common concepts and terminologies have been defined. At the moment, the requirements of the annual collection of national statistics for the National Institute for Health and Welfare are the only common factor between different client information systems [2]. Otherwise, the client data model and client documents of the social welfare are system-specific or municipality-specific. The data contents, like document structures, are heterogeneous at both the structure level and the semantic level.

Social workers need information from both information system of social welfare and from other data repositories. For example, the clients of social welfare are often also patients of health care. Most municipal social services have access to the query system of the Social Insurance Institution [3] and the population register system of the Population Register Centre [4]. In addition to this, there is hardly any exchange of information between the information systems of the social welfare. For example, in a survey on the use of information and communication technology in Finnish social services, only one of the five public service providers reported that digital exchange of information is possible [2].

The National Project for IT in Social Services (Tikesos) [5] initiated in 2004 and completed at the end of the year 2012. The main aim of the project was to improve the interoperability of client data in diverse information systems and to harmonize client data contents, semantics and structures. In the Tikesos project, the central information capital of the social welfare has been standardized so that the national archive being planned could be utilized to the management of client data. Separate information systems have to be able to send, receive, and process data so that the significance of the information will be preserved. The preservation of the significance of the information requires that the data structures have been standardized.

In this paper, our objective is to report how we can help information systems to utilize the RDFa-annotated XHTML documents. Fig. 1 demonstrates the forming of social welfare documents.

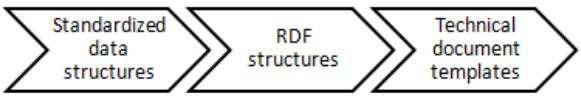

Fig. 1. Forming of Social Welfare documents

The Core Component Technical Specification (CCTS) [6] is used to form the data model of the social welfare (i.e. standardized data structures). The Resource Description Framework (RDF) [7] repository is used to store the standardized data structures and to query data in the RDF format. Finally, we generated RDFa-annotated XHTML documents (called technical document templates) from the RDF repository. The ontology which is used for the annotation is CCTS-based [6]. The technical document templates are the eXtensible HyperText Markup Language (XHTML) [9] documents where the Resource Description Framework – in – attributes (RDFa) [10] attributes are is used for embedding exchangeable information in the RDF format within XHTML documents.

The paper is organized as follows: in Section II we introduce how we specified CCTS-based data structures, in Section III we introduce how we formed the RDF repository which supports the CCTS-based data structures, in Section IV we present semantic documents (XHTML+RDFa) and illustrated utilizing of them. In Section V, we discuss in more details semantic assets of social welfare. Finally, the conclusions are given in Section VI.

II. Data Structures

Data structures, i.e. the data components (or classes) and document structures, together form the data model of the social welfare. The data components are semantic units which are used to structure the contents of the document. The data component consists of fields which can be based on the other data component or can be atomic fields. Fig. 2 shows an example of the data component `Private Person`.

Private Person
Last name : Name [1..1]
First names : Name [1..1]
Identity number : Identifier [0..1]
Temporary identity number : Identifier [0..1]
Date of birth : Date [0..1]
Birth information : Birth information [0..1]
Nationality : Code [0..*]
Contact information : Contact information [0..1]
. . .

Fig. 2. Example of data components

There are labels of the fields (e.g. `Last name`) in the class. The type of the field can be a simple information type (e.g. `Name`) [11] or it can refer to a class (e.g. `Birth information`). The obligatoriness (e.g. obligatory `[1..]`) and recurrence (e.g. only one `[..1]`) of the field are inside the square brackets.

Document structures consist of document-specific components which are based on data components. In addition, document structures may contain document-specific fields. The document-specific components and fields are modeled in structural form (Fig. 3).

Rectification

Header *[1..1]*
Client *[1..*] Demand for rectification applies to this person.*
Authority for appeal *[1..1] Authority to which appeal is appointed to.*
Receiver *[1..1] Receiver of demand for rectification*
Draftsman *[1..1]*
Basic informations of document *[1..1]*

	Definition
Content of demand for rectification *Text [1..1]*	Description from the sections of the original decision to which the client is dissatisfied or from the faulty decision and identified appeal demands.
Arguments *Text [1..1]*	Reasons why the client complains and arguments from the sections of law.

Fig. 3. Example of document structure

The name (e.g. `Client`), occurrence (e.g. `[1..1]`) and definition (e.g. `Demand for . . .`) of the document-specific components are shown. The name (e.g. `Arguments`), occurrence (e.g. `[1..1]`) and definition (e.g. `Fact reasons why . . .`) of the document-specific fields are shown. The type of the document-specific field is a simple information type (e.g. `Text`).

The document-specific components are specified. An example of the document structure (Fig. 4) visualizes the document-specific component `Client` which is a specification of the data component `Private Person`. When the document-specific component is specified, the necessary fields will be chosen from the data component.

Rectification

Header *[1..1]*
Client *[1..*] Demand for rectification applies to this person.*

	Definition
Last name *Name [1..1]*	Name that is shared by people in a family, e.g. Smith.
First names *Name [1..1]*	Name that is used as an additional name for the last name. The first name does not refer to the family. A person may have one or more first names, for example Matti Tapio. In applications the label of the field is in plural form "First names".
Identity number *Identifier [0..1]*	Finnish unique identifier of private person given by Population Register Centre
Temporary identity number *Identifier [0..1]*	Temporary finnish unique identifier of private person
Contact information *[0..1]*	*Contact information of private person*

Authority for appeal *[1..1] Authority to which appeal is appointed.*
Receiver *[1..1] Receiver of demand for rectification*
Draftsman *[1..1]*
Basic informations of document *[1..1]*

	Definition
Content of demand for rectification *Text [1..1]*	Description from the sections of the original decision to which the client is dissatisfied or from the faulty decision and identified appeal demands.
Arguments *Text [1..1]*	Reasons why the client complains and arguments from the sections of law.

Fig. 4. Example of specified document structure

We adapted the Core Components Technical Specification (CCTS) for describing the data components and document structures. CCTS can be applied to definition, storing, using and sharing of data [6]. It is suitable for defining data models and for creating standards for data exchange between organizations in an open, global environment. During the development process of the CCTS-based data model, the substance experts modeled the data components as well as document structures as spreadsheet tables because it was familiar form to them.

III. RDF STRUCTURES

We selected the Resource Description Framework (RDF) for storing and utilizing the data model because it is the most common approach [12] for linking data to ontologies which are used to improve data access. RDF is a standard model for data interchange with features that facilitate data merging. It specifies the conceptual data model of the information. However, RDF does not describe the semantics of data contents. Therefore, RDFS (RDF Schema) [13] and OWL (Web Ontology Lanquage) [14] can be used for adding semantics to the RDF data model. Fig. 5 illustrates the process of developing and applying the RDF repository. We developed a conversion tool for the generation of RDF structures through intermediate eXtensible Markup Language (XML) [15] structures from CCTS-based tables. Transformations were implemented by Python [16] and XSLT (Extensible Stylesheet Language Transformations) [17] scripts.

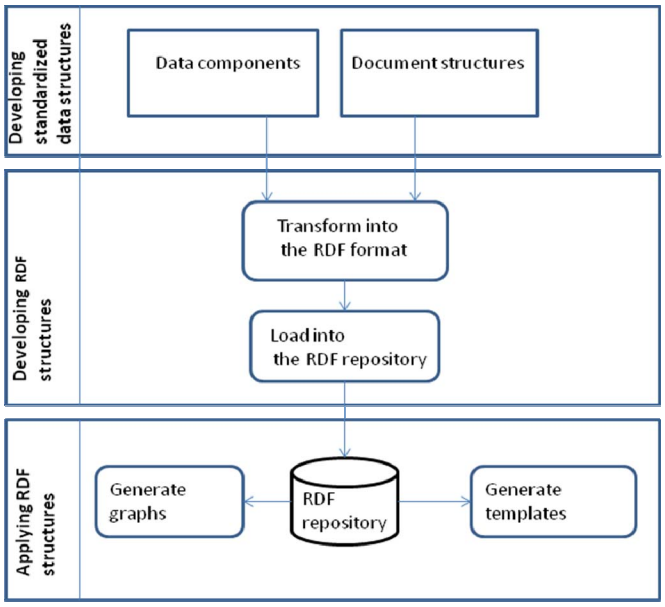

Fig. 5. Forming of RDF repository

The ontology with linked data was stored to the RDF repository which allows to link data together with URIs (Universal Resource Identifiers) [18]. We chose the Sesame open source Java framework [19] for storing data components and document structures and querying data in the RDF format. Sesame is deployed as a Java Servlet Application in Apache Tomcat webserver whereby client applications communicate over HTTP. XSLT+SPARQL [20] is used for making queries and generating result HTML or XHTML pages. The SPARQL Query Language for RDF (SPARQL) [21] queries are executed against RDF repositories, and the results can be accessed using XPath [22]. In this way, it is possible to write XSLT transformations which access data that is expressed in RDF.

IV. TECHNICAL DOCUMENT TEMPLATES

Technical document templates are automatically generated document structures in a XHTML+RDFa format. System suppliers can utilize templates in their information systems. The XSLT+SPARQL script generates the XHTML documents where the RDFa recommendation is used for embedding exchangeable information in the RDF format within XHTML documents.

For making the versioning of the archived documents possible, the document-specific information must be located in an individual XHTML+RDFa file. The document contains a set of Cascading Style Sheets (CSS) [23] rules for default formatting. It is also possible to replace or extend this default set of rules for tailoring the default formatting.

The prefix attribute of the element body defines the namespace of the document. URI of the data structure is constructed by adding the technical name of a data structure (e.g. *ak:ClientPrivateperson*) to the end of the document namespace. The technical document template has a structure which follows the CCTS-based structure of the RDF data. Furthermore, comments remark if the data structure is obligatory or repeatable. The following fragment of the document is an example of the technical document template.

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+ RDFa
1.1//EN" "http://www.w3.org/MarkUp/DTD/xhtml-
rdfa-2.dtd">
<html>
  <head>
    <title>Rectification</title>
    <meta http-equiv="Content-Type"
          content="text/html; charset=UTF-8"/>
    <style type="text/css">
       div          { margin: 0; . . .}
       div.document { margin-top: 10mm; . . .}
       . . .
    </style>
  </head>
  <body
   prefix="ak:http://sosmeta.fi/asiakirjat/
    appeals_Clientdocumentation/Rectification/
    2012/08/28/ak#"
   typeof="ak:Rectification">

<!—Document begins-->
    <div class="document">
       . . .
       <div class="bdtitle">
            Rectification
       </div>
<!—Data component begins: Client-->
<!--Definition: Demand for rectification
applies to this person.-->
<!--This reference is obligatory!-->
<!--This reference can be repeated!-->
    <div class="item"
        rel="ak:ClientPrivateperson">
        <div class="title">
             <b>Client</b>
        </div>
    </div>
<!--Data component ends: Client-->
      …
```

```
<!-- Document-specific field begins: Arguments
-->
<!--Definition: Fact reasons why client wants
to have a change and the sections of law to
which the client appeals in.-->
<!--This reference is obligatory!-->
   <div class="item">
       <div class="title">Arguments</div>
         <div class="field"
            property="ak:ArgumentsText">
             The client did not understand
             that she has to deliver the
             needed appendixes.
         </div>
      </div>
   </div>
<!—Document-specific field ends: Arguments-->
    …
  </div>
<!—Document ends-->
 </body>
</html>
```

Data from information systems is added as a value of the div elements with RDFa attributes property or rel (e.g. `property="ak:ArgumentsText"`). The user of the technical document template can remove a part of the structure if it is not compulsory according to the client data model. Part of the described structure can also be repeated in the document depending on the client data.

The main stages of the utilizing of technical document templates are the following:

1) Determine the correspondence of the information definitions of the data model of the client information system and of the data model of the social welfare. The correspondence of information can be examined with the help of the document structures and of the generated graphs.

2) Form the structure of the database which is in accordance with the data model of the social welfare.

3) Connect the fields (the RDFa attributes) that have been used in the document template to the client data of the client information system.

4) Check that the documents which are in accordance with the document template are valid according to the validation service (the formed document examples must be checked with the help of validation services for example W3C offers the XHTML validation service [24] and CSS validation service [25]).

The technical document templates can be utilized in the information systems of the social welfare in many different ways as follows:

- The XHTML+RDFa structure of the document can be treated for example with different programming languages which support the Document Object Model (DOM) [26] defined by the W3C.
- The technical document template can be used as a basis for the information system document templates (for example jsp, scala or php template) which are used to add client data to the technical document templates.
- The document which has been completed with the client data can be generated by XSLT or XQuery [27] from the XML format which is specified by some client data system. In that case, the technical document templates can be used as a basis for the XHTML+RDFa document with client data.

V. DISCUSSION

Semantic assets are as instances of the know-how capital. Usually, the definitions of the semantic assets or semantic interoperability assets are mainly asset lists as follows:

- ”dictionaries, taxonomies, mapping tables, XML schemas . . . ontologies, tags, ontology-folksonomy maps” [28]
- “ontologies, data models, data dictionaries, code lists, XML (Extensible Markup Language) and RDF (Resource Description Framework) schemas which are used for information exchange and that can be reused by implementers of Information Systems, in particular, as part of machine-to-machine interfaces” [29]
- “dictionaries, thesauri, taxonomies, mapping-tables, ontologies and service registries”, “highly reusable metadata (e.g. xml schemata, generic data models) and reference data (e.g. codelists, taxonomies, dictionaries, vocabularies) which are used for eGovernment system development” [30]
- “the resources required to enable semantic interoperability such as terminologies, thesauri and mapping rules” [31]

Versatile know-how (e.g. substance, terminological and technical) is required for the interpretation of the semantic assets, not to mention, for the understanding and adapting of the semantic assets. Therefore, we reported on the concrete solution to forming and adapting the semantic assets of social welfare (Fig. 6).

Fig. 6. Semantic Assets of Social Welfare

We wanted to use the RDF repository for storing and utilizing the data model. In our review of the related work, we used Google Scholar as a pilot data source and we tried search terms RDF and CCTS and RDFa and XHTML from titles. However, we did not find out any RDF paper either for CCTS-based storing or RDFa-annotated XHTML documents utilizing. There is, for example, Europeana RDF Store Report [32] which summarizes the results of qualitative and quantitative study which were carried out on existing RDF

stores in the context of the European Digital Library project. Peristeras, Tarabanis and Goudos [33] have examined the projects, models and ontologies of the electronic administration (eGovernment). They have classified and grouped 29 different eGovernment modeling initiatives. Two of those used the RDF standard; The UK Government Common Information Model (GCIM) [34] and The Federal Enterprise Architecture (FEA) Business Reference Model [35].

VI. CONCLUSION

The standardized data structured of the social welfare in Finland have been specified as spreadsheet tables because they are familiar forms to the substance experts. The data components and document structures form the standardized data structures (i.e. CCTS-based data model). Data structures have the typed fields that can be, for example, obligatory and recurrent.

The RDF repository has been chosen for utilizing the CCTS-based data model. The Python programming language and XSLT language have been used for forming the RDF structures. Two languages, XSLT and SPARQL, have been used for generating graphs and templates from the RDF repository. The technical document templates are examples of generated templates. They are RDFa-annotated XHTML documents and they have been formed to help utilizing the semantic documents of the social welfare. All spreadsheet tables, RDF structures, as well as, generated graphs and templates are semantic assets of the social welfare.

The data model of the social welfare will be estimated with the inspection during the years 2013–2015. In the inspection, the necessity, adequacy and purpose of use of documents, among others, are estimated. Also the contents of documents, such as the necessity, adequacy, notation, obligatoriness and recurrence of the information are estimated. Inspectors have to be able to read and interpret the data model. Thus, the inspectors have to understand how the data model has been formed.

In the future, we have to follow several issues (e.g. standards) around the semantic documents. Furthermore, we have to develop processes for forming and handling of the semantic documents in the information systems. Therefore, we have specified business rules and prototyped those with Drools [36].

REFERENCES

[1] H. Eriksson, "The semantic-document approach to combining documents and ontologies," Int. J. Human-Computer Studies 65, pp. 624–639, 2007.

[2] P. Hämäläinen, J. Reponen, I. Winblad, J. Kärki, M. Laaksonen, H. Hyppönen, M. Kangas, "eHealth and eWelfare of Finland - Checkpoint 2011," National Institute for Health and Welfare - Report 5/2013. http://www.julkari.fi/bitstream/handle/10024/104368/URN_ISBN_978-952-245-835-3.pdf?sequence=1

[3] The Social Insurance Institution of Finland. http://www.kela.fi/in/internet/english.nsf

[4] The Population Register Centre. http://www.vrk.fi/default.aspx?site=4

[5] Sosiaalialan tietoteknologiahanke. http://www.sosiaaliportti.fi/fi-FI/tikesos/in-english/

[6] United Nations Centre for Trade Facilitation and Electronic Business. Core Components Technical Specification, Version 3.0, 2009. http://www.unece.org/fileadmin/DAM/cefact/codesfortrade/CCTS/CCTS-Version3.pdf

[7] W3C, Resource Description Framework (RDF). http://www.w3.org/RDF/

[8] K. Hyppönen, M. Alonen, S. Korhonen, V. Hotti, "XHTML with RDFa as a Semantic Document Format for CCTS Modelled Documents and its Application for Social Services," The Semantic Web: ESWC 2011 Workshops, Lecture Notes in Computer Science, vol. 7117/2012, pp. 229-240, 2012.

[9] W3C, XHTML™ 1.1 - Module-based XHTML - Second Edition, W3C Recommendation 23, November 2010. http://www.w3.org/TR/2010/REC-xhtml11-20101123

[10] W3C, RDFa 1.1 Primer, Rich Structured Data Markup for Web Documents. http://www.w3.org/TR/2012/NOTE-rdfa-primer-20120607/

[11] The Advisory Committee on Information Management in Public Administration (JUHTA), JHS 170. http://www.jhs-suositukset.fi/web/guest/jhs/recommendations/170

[12] A. Hertel, J. Broekstra, H. Stuckenschmidt, "RDF Storage and Retrieval Systems," in Staab, Studer, (eds), Handbook on Ontologies, second edition, pp. 1-17, Springer, 2009.

[13] W3C, RDF Vocabulary Description Language 1.0: RDF Schema,W3C Recommendation, 10 February 2004. http://www.w3.org/TR/rdf-schema/

[14] W3C, Web Ontology Language (OWL). http://www.w3.org/2004/OWL/

[15] W3C, Extensible Markup Language (XML). http://ww.w3.org/XML/

[16] Python Programming Language, Python Software Foundation. http://www.python.org/

[17] W3C, XSL Transformations (XSLT) Version 2.0, W3C Recommendation 23 January 2007. http://www.w3.org/TR/xslt20/

[18] Network Working Group. Request for Comments: 2396, Uniform Resource Identifiers (URI): Generic Syntax, August 1998. http://tools.ietf.org/html/rfc2396

[19] OpenRDF.org. http://www.openrdf.org/

[20] D. Berrueta, J.E. Labra, I. Herman, "XSLT+SPARQL: Scripting the Semantic Web with SPARQL embedded into XSLT stylesheets," Proceedings of 4th Workshop on Scripting for the Semantic Web, Tenerife, Jun 2008. http://berrueta.net/file_download/5

[21] W3C, SPARQL Query Language for RDF, W3C Recommendation, 15 January 2008. http://www.w3.org/TR/rdf-sparql-query/

[22] W3C, XML Path Language (XPath), Version 1.0, W3C Recommendation 16 November 1999. http://www.w3.org/TR/xpath/

[23] W3C, Cascading Style Sheets Level 2 Revision 1 (CSS 2.1) Specification, W3C Recommendation 07, June 2011. http://www.w3.org/TR/CSS2/

[24] W3C, Markup Validation Service. http://validator.w3.org/

[25] W3C, CSS Validation Service. http://jigsaw.w3.org/css-validator/

[26] W3C, Document Object Model (DOM). http://www.w3.org/DOM/

[27] W3C, XQuery 1.0: An XML Query Language (Second Edition), W3C Recommendation 14 December 2010. http://www.w3.org/TR/xquery/

[28] A. Ojo, E. Estevez, T. Janowski, "Semantic interoperability architecture for Governance 2.0," Information Polity - Government 2.0: Making Connections between citizens, data and government, vol. 15, issue 1,2, pp. 105-123, April 2010.

[29] J.R. Frade, D. Di Giacomo, S. Goedertier, N. Loutas, V. Peristeras, "Building semantic interoperability through the federation of semantic asset repositories," Proceedings of the 8th International Conference on Semantic Systems (I-SEMANTICS '12), Harald Sack and Tassilo Pellegrini (Eds.), ACM, New York, NY, USA, pp. 185-188, 2012. http://doi.acm.org/10.1145/2362499.2362528

[30] Interoperability Solutions for European Public Administrations (ISA), Towards Open Government Metadata. https://joinup.ec.europa.eu/sites/default/files/towards_open_government_metadata_0.pdf

[31] European Interoperability Framework (EIF) for European public services. http://ec.europa.eu/isa/documents/isa_annex_ii_eif_en.pdf

[32] B. Haslhofer, E. Momeni, B. Schandl, S. Zander, “Europeana RDF Store Report”, 2011. http://eprints.cs.univie.ac.at/2833

[33] V. Peristeras, K. Tarabanis, S.K. Goudos, “Model-driven eGovernment interoperability: A review of the state of the art,” Computer Standards & interfaces, vol. 31, issue 4, pp. 613-628, 2009.

[34] The UK Government Common Information Model (GCIM), Office of e-Envoy, e-Services development framework primer, UK cabinet Office, 2002.

[35] The Federal Enterprise Architecture (FEA) Business Reference Model, CIO Council, Federal Architecture Enterprise Framework v.1.1, 1999.

[36] Drools - The Business Logic integration Platform. http://www.jboss.org/drools/