Title: Using XML for Long-term Preservation

Subtitle: Experiences from the DiVA Project

Authors: Müller, Eva; Klosa, Uwe; Hansson, Peter; Andersson, Stefan; Siira, Erik

Organization: Uppsala University Library, Electronic Publishing Centre
Email: eva.muller@ub.uu.se; uwe.klosa@ub.uu.se; peter.hansson@ub.uu.se; stefan.andersson@ub.uu.se; erik.siira@ub.uu.se

Address: Box 510, 75 120 Uppsala, Sweden

Url: http://publications.uu.se
Keyword: long-term preservation, XML, XML Schema, DiVA, DiVA Document Format, DiVA Archive, URN, URN:NBN

Abstract

One of the objectives of the DiVA project is to explore the possibility of using XML as a format for long-term preservation. For this reason, the practical use of XML in different parts of the system was evaluated before deciding on the design.

The DiVA Document Format - defined by an XML schema - has been developed to describe the inter-relationships amongst the various data elements and processes, and to support long-term preservation of the actual documents.

XML Schema provides a means for defining the structure, content and semantics of XML documents. It is an XML based alternative to the XML Document Type Definition (DTD). Because one of the primary reasons for using XML was to support long-term preservation, the most popular DTDs for documents: DocBook and TEI were evaluated. Limitations regarding metadata descriptions were found in both of these DTDs, so the decision to develop a new structure for DiVA, using XML schema, was made. This schema combines the DocBook Schema (derived from the DocBook DTD) for the textual parts of the document with the internal schema for all metadata (bibliographic and administrative data).

Using the DiVA Document Format for content management and inter-process communication, several applications were developed. Some of their purposes are essential for long-term preservation:

- Make persistent National Bibliographic Numbers (NBN) available for the URN resolution service[1] at the Royal Library in Stockholm available.
- Send MARC21 records in MARC-XML to the National Library.
- Create archival file packages for long-term preservation, checksum them, store them in the DiVA Archive and send a copy of them to the Swedish Royal Library.

Currently the file-archives for long-term preservation contain the original full-text file in various formats and the DiVA Document Format file, which contains all the metadata about the document. Furthermore the DiVA Document Format file contains all parts of the full-text file that can be converted into XML. In the future it might be possible to transfer the whole full-text into XML, in which case the file-archives would contain only DiVA Document Format files.

---

[1] http://urn.kb.se/resolve

Table of Contents

Preface

DiVA - Digitala vetenskapliga arkivet (DiVA Archive) - is a comprehensive description of a searchable archive containing all documents, which are published in an electronic form at Uppsala University in Sweden. Other Swedish universities are also co-operating in the project within the DiVA framework. One part of this archive is the database containing theses published at Uppsala University from 1998 to date.

In September 2000 an Electronic Publishing Centre was established at Uppsala University Library. Its primary assignment was a project in which technical solutions, and a well-functioning workflow, for electronic posting and full-text publication of doctoral theses, essays, working papers and other types of scientific publications were to be created.

The first phase of the project was completed in 2002 and the result was the DiVA Publishing System – a system for electronic publishing of different types of publications.

One of the goals has been to create a long-term archive containing all digital documents published at Uppsala University. The assignment involves both technical and organisational issues. Developer team faced with many questions. How can the loss of data be avoided? What kind of descriptive and administrative metadata is useful for archiving? What is the appropriate metadata format for long time preservation? How important is the layout of the objects and how is it to be handled? How can images and formulas be handled?

Because of those questions, XML was discussed early on as a format for storing descriptive and administrative metadata, as well as for the complete content of the documents. XML represents a format that is easy to restore and understand by both humans and machines.

This paper will describe the current status of the XML implementation in DiVA Archive and the surrounding applications and why XML is an important format for long-term preservation.

1 XML as Long-term Preservation Format

One of the objectives of the DiVA project is to explore the possibility of using XML as a format for long-term archiving.

There are several advantages of using XML encoded documents for long-term archiving. XML is an open and established notation. XML documents are in a human-readable text format and internationalised character sets are supported. These characteristics facilitate data migration and the documents are likely to have longevity. For these reasons XML seemed like a good choice, but to ensure success, the practical use of XML in different parts of the system was evaluated before a decision about the design was made.

In the DiVA project XML is not only for archiving. It is also used for the communication between different processes within the system and for the internal communication in the development team. It also helps to validate data with help of an XML schema. The dynamic web interface is built on XML and XSLT.

1.1 XML Schema

XML Schema provides a means for defining the structure, content and semantics of XML documents. XML Schema is an XML based alternative to the XML Document Type Definition (DTD). Because the primary reason for using XML was to support long-term archiving, the most popular DTDs and schemas for documents namely DocBook and TEI were evaluated. Limitations regarding the metadata descriptions needed in the DiVA project were found.

Because of the need to combine administrative metadata, descriptive metadata and content, a new schema was developed that meets the needs of the DiVA project. This schema combines the DocBook schema (derived from the DocBook DTD) for the textual parts of the document with the bibliographic metadata and administrative metadata for long-term preservation.

XML Schema was chosen over XML DTD because it is written in XML and supports many data types, self-defined data types and different namespaces. The support for different data types offers several advantages. It is possible to describe permissible document content, to validate the correctness of data, to define restrictions on data (data facets), to define data formats (data patterns) and to convert between different data types. It is also easier to work with data coming from a database.

During the development, it was noticed that XML Schema facilitated the communication between the developers by providing a simple mechanism for writing formal specifications of subsystem interfaces.

1.2 Comparison of DocBook and TEI

TEI[2] and DocBook[3] are two widely used recommendations for encoding textual material in electronic form. These two recommendations were compared to find which is most appropriate and convenient to use when representing full-text documents in the DiVA Archive.

A logical unit, i.e. a combination of XML elements and/or XML attributes that have a certain well-defined meaning, can be expressed differently in TEI and DocBook. A logical unit that consists of only one well-defined element in DocBook often is composed by both a general element and attribute in the TEI representation. Attribute values are not defined in the TEI recommendation and therefore have to be defined locally. Therefore it is likely that others would not correctly interpret a TEI encoded document without any agreements.

Elements that define the structure of documents, e.g. headers, chapters, lists and tables are more

---

[2] See: http://www.tei-c.org/

[3] See: http://www.docbook.org/

specifically defined in DocBook than in TEI. For publication of documents like PhD theses or scientific papers it is therefore more convenient to use DocBook because relevant structure elements are well defined. But if a text should be marked-up in detail both semantically and structurally, for example in order to create scholarly archives of diverse kinds of historical sources or for linguistic purposes, the more general TEI scheme would be a better choice.

The main purpose in the DiVA project is to store the structure of the contents of the documents and not to store the semantics. Therefore DocBook was chosen to mark up the content.

Element	TEI	DocBook
Heading 1	<div1 type="chapter" n='1'> <head n="1">Heading 1</head> </div1>	<chapter id="1"> <title> Heading 1</title> </chapter>
Superscript	<hi rend="sup">text</hi>	<superscript>text</superscript>
Lists	<list type="…"></list>	<orderdlist numeration="…">…</orderdlist>

Table 1: Some elements in TEI and DocBook

1.3 DiVA Document Format

DiVA Document Format - defined by an XML Schema - version 1.0 consists of 99 elements[4]. Administrative elements are combined with descriptive elements to make it possible to describe a publication in the same XML document file that contains its content. Many element names exist in both singular and plural form. The plural form is always used to name container elements. A container element contains one or more elements in its corresponding singular form. For example <creators> contains one or more <creator> elements, <titles> contains <title> elements and so on. The container elements group elements that contain the same type of information. These container elements can also group elements that contain closely related information[5]. This makes it easier for human readers to find information quickly in the document. Machines can also benefit from the fact that the distance between interrelated information is kept short.

One of the advantages of XML Schema over DTD is that it has many in-built data types such as numerical values and dates. When applicable the predefined XML Schema data types have been used in the DiVA Document Format. But there are exceptions when the built-in types are not an appropriate choice. An example is the element xs:date. xs:date represents a date as defined in ISO 8601. The lexical form is CCYY-MM-DD. Since specifying xs:date would require a year (CCYY), month (MM) and date (DD), it was necessary to define a different date format that would allow one to specify the year, even when month and date are unknown.

XML Schema supports user-defined complex types. A complex type describes complex structures built by elements and other types. The most commonly used complex types in DiVA Document Format are personType (see appendix Figure 2) and organisationType (see appendix Figure 3) which define a person and an organisation respectively.

All XML elements defined in the XML Schema have English-language names. There are both general elements and specific elements defined by the XML Schema. General elements facilitate introducing new concepts to the XML Schema without changing it (promote scalability of concepts). In spite of this fact it is not always suitable to generalize well-defined concepts too much as in the case of creators and contributors. An option would have been to use a person or an organisation element with attributes. The documents should also be easy to read for humans and therefore both general and specific elements have been defined.

---

[4] See: http://publications.uu.se/schema/1.0/diva.xsd

[5] This can be elements which only exist for a specific type of documents.

The order of the elements was not the focus while developing the DiVA Document Format, though, there were some exceptions. The <properties> element has to be the first child element if applicable. A property should be near the parent element it describes. The child elements <specifics> have to be in ascending order of generality, i.e. the specific groups have to come before the more general groups of elements. In object-oriented programming languages, this facilitates the creation of raw XML from objects in such a way that the transformation process can start to translate specific information stored in the subclasses and end up translating the general data in the super classes.

DiVA Document Format defines one root element, which is called <documents>. This makes it possible to save more than one document in one XML file, which is needed in some applications in the DiVA project[6]. But for archiving purposes each document is stored in a separate file containing a <documents> element with exact one <document> element.

The <properties>-element (and its child <property>) is used in several constructs and is used to give the parent element arbitrary attributes (not to be confused with XML attributes). Each property is defined independently of the others. If dependencies are crucial, another construct has to be used. An XML attribute is sometimes used on the property element to make it clear what the property stands for. This construct has several advantages: new properties can be easily integrated into the schema without changing the main structure of the definition (scalable), every level in an hierarchical structure can be described in a plain fashion as properties, and when a property has an XML attribute it is easy for a machine to find the right section in the XML document.

<identifiers> is widely used in the documents conforming to DiVA Document Format. The <identifiers> element gives the parent element one or more identifiers. Each <identifier> element consists of a (<properties>,<value>)-pair. The properties describe the identifier and the identifier itself is specified under the <value> tag. The following identifiers are in use today: “local”, which is an identifier that is only used within an organisation; “internal”, which currently binds XML data to a relational database system; “ISSN” and “ISBN”, which are used to identify series and publications; “URI” (uniform resource identifier), formatted as a web link (URL); “URN:NBN”, a special national unique identifier; and country and language codes according to ISO639 and ISO3166, respectively.

The URN:NBN identifier is used to map electronic resources to URLs and as a primary key of the publications stored in DiVA.

<specifics> is a container element that contains child elements that are not generally applicable. It is convenient to put all elements that only exists for a certain publication type into a place where they can be found easily by both humans and machines.

<manifestations> is an element that contains different manifestations or instances of the same publication[7]. Because a document can be stored in many different formats, both physical and electronic, each of them must be described individually. The formats can, for example, have different identifiers, be member of different series, be published and distributed at certain dates by different organisations and/or persons. Today most of the doctoral theses stored in the DiVA system have two manifestations, a physical book and an electronic PDF file.

The manifestation element contains also metadata about migration from one format to another format. If, for example, a PDF manifestation was migrated to a newer version of PDF, a new manifestation is created with information about the original manifestation, which is stored in the archive, too.

2 Long-term Preservation in the DiVA Project

Five Swedish universities cooperate within the DiVA project, which originated at the Uppsala University. The participants are the universities of Stockholm, Södertörn, Umeå, Uppsala and

---

[6] Delivery of search results to the search interface on the website and the application to maintain the archive.

[7] Some inspiration has been gathered from the work concerning Functional Requirements for Bibliographic Records, FRBR, by IFLA. See: http://www.ifla.org/VII/s13/frbr/frbr.htm

Örebro. The main goals of the project are to create a searchable archive for long-term preservation and to disseminate the scientific work of the five universities

Long-term preservation is only useful, if several copies of the archive exist and a persistent and unique identifier identifies every document. Therefore several projects were initiated in cooperation with the Royal Library in Stockholm. At the Royal Library in Stockholm National Bibliographic Numbers (NBN, see section 2.1) are made available to a URN resolution service[8] at the Royal Library in Stockholm. MARC21 records in MARC-XML are sent to the National Library (LIBRIS). These records contain the URN:NBN of the described document. The catalogue at the National Library and the archives at the Royal Library are likely to be well maintained and have longevity, so they are relatively safe places to deposit documents for long-term preservation.

And since the main purpose of the project initiated with the Royal Library is to create a copy of the local DiVA Archives[9], check-summed file-packages are sent there (see section 2.2).

2.1 Uniform Resource Name (URN) and National Bibliographic Number (NBN)

A uniform resource name, or URN, is a unique and permanent identifier for electronic resources on the Internet. Unlike a uniform resource locator, i.e. an URL, an URN is a permanent identifier that cannot be changed over time. An URN cannot be assigned to other resources even if the mapped resource has ceased to exist. The national library of Sweden assigns URNs in the Swedish national bibliographic number domain (URN:NBN:se) to organisations and the public in Sweden.

The DiVA archive has been assigned the sub domains URN:NBN:se:X:diva where X stands for an abbreviated form of the participant in the project. Uppsala University is abbreviated uu (umu for Umeå University, oru for Örebro University, sh for Södertörn University, su for Stockholm University). To automatically give every published document in the DiVA archive an unique URN:NBN identifier a serial number is added to the sub domain, e.g. URN:NBN:uu:diva-3100.

To resolve a URN:NBN a resolution service has been developed and installed at the Royal Library in Stockholm. To resolve the example above the following URL can be used http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-3100. The Royal Library has published guiding principles for the use of URN:NBN in Sweden at http://www.kb.se/urn/riktlinjer.htm[10].

In case of the closing of the DiVA archive at Uppsala University, the respective URN:NBN will be resolved to the National Library Archive. In this term the long-term preservation copy is directly connected to the resolution service.

2.2 The DiVA Archive

Today the DiVA Archive consists of metadata files conforming to the DiVA Document Format, full-text files in voluntary formats (PDF is the most common) and checksum files. The files are stored in dedicated folders in an ordinary file system. The files have controlled names building on the URN:NBN identifiers. The identifiers link metadata files to theirs respective full-text files. Our experience so far shows this is a convenient way to store data.

The archive has been developed according to the OAIS (Open Archival Information System) framework and reference model[11]. The administrative metadata in the DiVA Document Format refer to the OAIS model.

The structure of the archive is hierarchical and can be easily mapped to a file system or a native XML database. The root directory is called *$archive_home* and must be specified in the

---

[8] This link can be used for resolving an URN:NBN: http://urn.kb.se/resolve?urn=...

[9] See the agreement: http://www.ub.uu.se/diverse/avtal.pdf (in Swedish)

[10] There is also a RFC published at http://www.ietf.org/rfc/rfc3188.txt

[11] See: http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

environment the archive system runs on. There is a file called *readme* in the root folder that describes the archive. The *readme* file contains information about the archive's structure, file name conventions, description of concepts, checksum algorithms[12] and references to documents stored in the archive that contain important information about the archive.

Each document in the archive has been assigned a unique and permanent URN:NBN identifier. The identifiers' constituent parts, which are separated by colons or hyphens, build up the hierarchical structure of the archive. For example the document with the identifier urn:nbn:se:uu:diva-1144 gives rise to the folder *$archive_home*/urn/nbn/se/uu/diva/1144 in the archive. This folder is also called a document root folder.

A document root folder contains a changeable metadata file and numbered folders, starting with 1, containing different manifestations of the same document. The changeable metadata file can be altered any time, both before and after manifestations are published. The metadata file conforms to the DiVA Document Format (see section 1.3).

The name of a manifestation folder refers to a number specified in the manifestation section of the metadata file. In addition to one or more full-text files a manifestation folder contains a copy of the changeable metadata file located in the document root folder. The copy is made automatically when a manifestation of the document is published. The metadata copy cannot be changed after the manifestation is published and therefore as opposed to changeable metadata is called unchangeable metadata. A manifestation folder can also include supplementary files, e.g. errata, and files (often images) that for example the full-text files link to.

Every file in the archive gets a checksum. A checksum is a number, calculated from the contents of a file, which is used to determine if the contents of a file are correct (i.e. to check a file's integrity). In the DiVA Archive checksums are stored in separate files.

The files stored in the archive are also described in the metadata file. During the project it was noticed that filename guidelines are likely to change over time. It is therefore essential that the file-properties, e.g. file formats, are stored in the metadata file together with the names of the files they describe. Properties described in the metadata file include filename, document type (full-text, supplement), file size, file format (abbreviation, version, identifiers to file format registry, description).

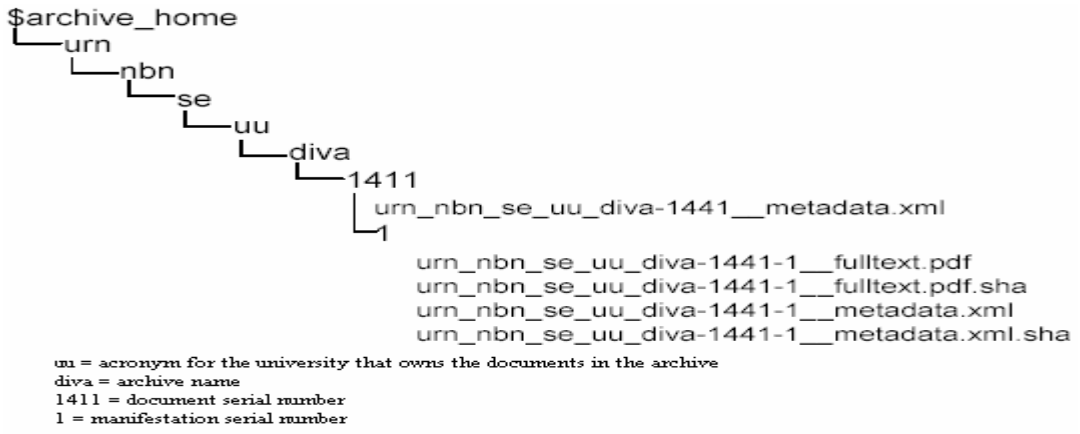

Figure 1: Structure of the DiVA Archive

After a certain period[13] from publishing, the document and its metadata are delivered as a package to the Royal Library for long-term preservation. This package contains only selected manifestations of the document, the DiVA Document Schema with all bibliographic metadata and administrative metadata. Each package is named after its URN:NBN identifier followed by the manifestation number it contains. This guarantees distinguishable and unique package names, e.g. urn:nbn:se:uu:diva-1144-2.

---

[12] The algorithms are SHA and MD5

[13] 6 months as an first assumption

3 Conclusions

The long-term preservation of digital objects includes a variety of challenges. The DiVA project has faced some of the technical problems concerning storage and storage media to guarantee the maintenance and the security of the DiVA Archive. But the focus of the DiVA project has been ensuring the future use and understanding of the digital objects in the archive can be assured. There is no guarantee that it will be possible to use and to understand these objects in the distant future, but there are ways to increase the chances likelihood of success.

This assumption was the starting point for the discussions about the design of the DiVA Archive and the DiVA Workflow. We tried to find a practical and convenient way to minimize risks for data loss, especially in the context of migration of the entire document and the connected metadata to other formats and media. Another important condition was to find a practical solution that is applicable for large-scale production and that can be part of an automated workflow. As a part of this workflow we established a connection to the National Library Archive, so that both the metadata of the digital objects and the digital objects themselves could be exchanged.

XML was discussed early on within the DiVA developer team as a possible format for long-term preservation. XML is an open and established notation. XML documents are in a human-readable text format and internationalised character sets are supported. These characteristics facilitate data migration and the documents are likely to have longevity. Therefore the decision was made to use XML as a format for storing descriptive and administrative metadata, as well as for the complete content of the digital objects. Thus, the DiVA Document Format was created.

The DiVA Document Format was developed to be compatible with a number of commonly used metadata standards relevant to electronic documents. At the present, however, we need to focus on the management of two other types of content - images and formulas. The integration of these into the DiVA Document Format is still under development. Formulas in abstracts are already stored as MathML, but the MathML must be created manually and inserted into the abstract[14]. For a production workflow it is absolutely necessary to have a tool that can construct these formulas automatically.

The latest results in our development are the transformation from MS Word documents to the DocBook format with the help of Open Office. Open Office is able to load MS Word documents and it stores documents using an internal XML. Our development team has created an XSL style sheet that transforms the Open Office XML into the DiVA Document Format. Open Office XML already stores formulas as MathML, so it may be possible, in the near, future to use Open Office to move formulas directly into the DiVA Document Format.

Because of the DiVA Document Format and the DiVA Archive, the first fundamental steps of the construction of an archive for long-term preservation have been taken. The usage of URN:NBN as an unique identifier and the exchange of metadata and archive-files with the National Library Archive, were the next important steps. Though the first implementation of the entire archiving workflow will not be completed until autumn of this year, electronic format will already be the primary mode of publication for some theses during next semester. This decision to move so quickly – made by faculty and the authors themselves – demonstrates their confidence in the DiVA Archive.

The outcome of the work, we have presented here, is not only the result of the work of our development team. It is also the result of discussions with other developers, librarians and researchers. We would especially thank our reference group[15] for their feedback and the support they continue to give us.

Tim DiLauro gave us useful feedback on this paper.

---

[14] With the help of the features in Open Office http://www.openoffice.org

[15] http://publications.uu.se/epcentre/diverse/refgrupp.html

Appendix

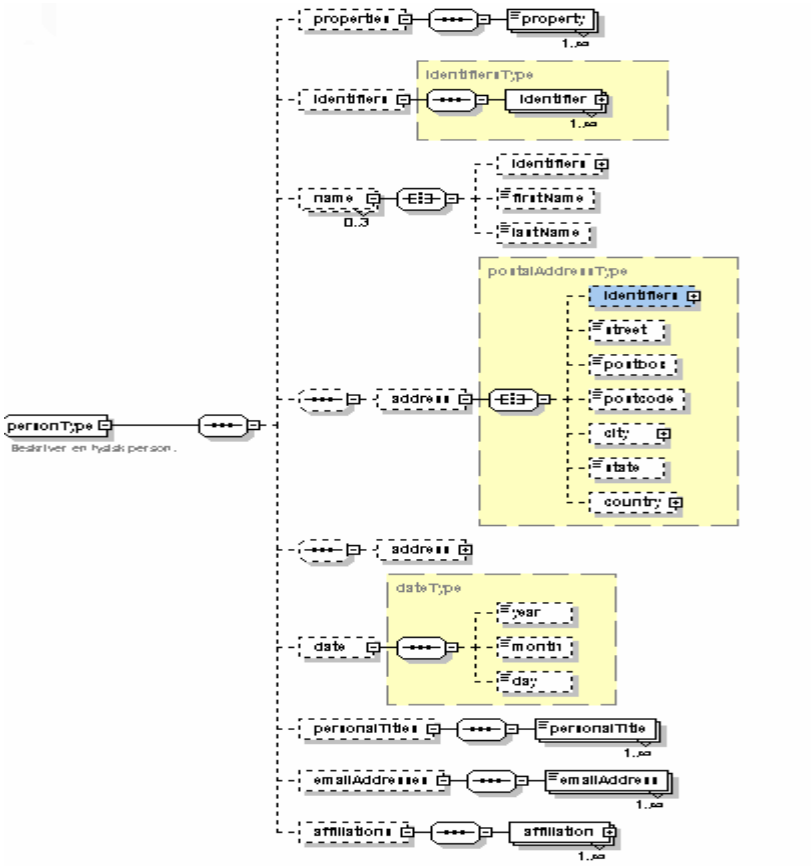

Figure 2: Graphical representation of the complex type personType

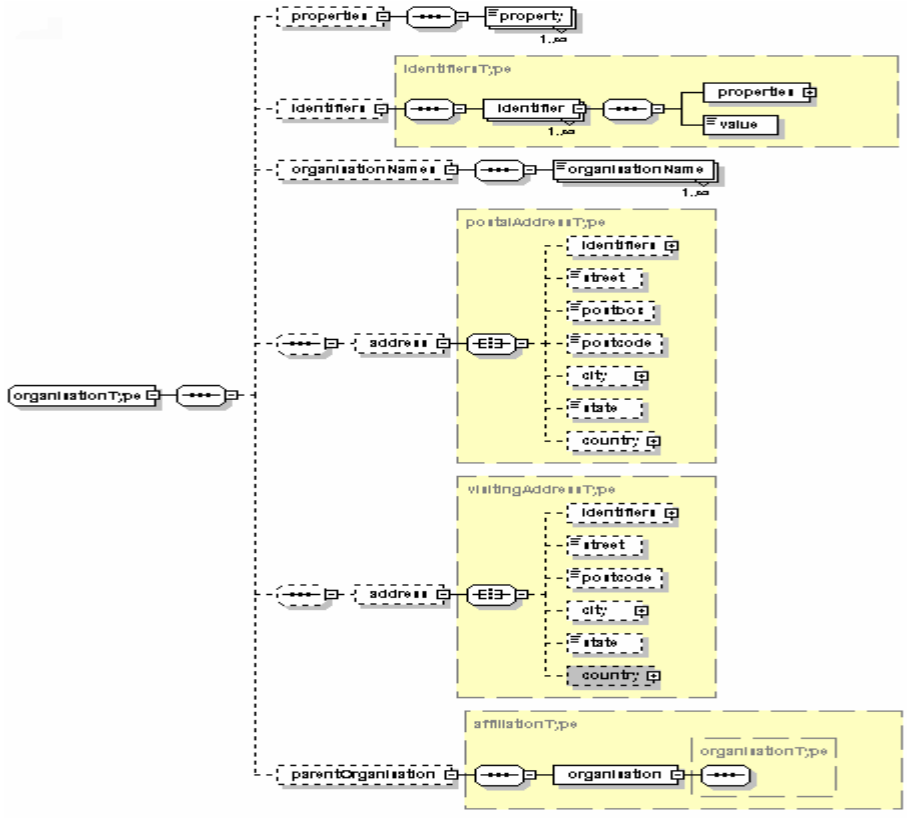

Figure 3: Graphical representation of the complex type organisationType