An Ontology for Portability and Interoperability Digital Documents
An approach in document engineering using ontologies

Erika Guetti Suca[1] and Flávio Soares Corrêa da Silva[1]
[1]*Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brasil*
{eguetti, fcs}@ime.usp.br

Keywords: Document Engineering, Document Interoperability, Document Portability, OOXML (Office Open XML), ODF (Open Document Format), Ontologies.

Abstract: Organizations need to exchange information simple and efficient, with costs as low as possible. Such information is ussually presented as documents with pre-defined content. These documents may be equivalent or almost equivalent but quite distinct in different organizations. The same document can be different depending on the historical context. Also, organizations do not always use the same technology to generate your documents. The purpose of this work is to enable interoperability of documents and achieve portability of digital documents through the reuse of content and format in different plausible combinations. We propose the characterization of digital documents using ontologies as a solution to the problem of lack of interoperability in the implementations of document formats. As proof of concept we consider the portability between OOXML and ODF document formats.

1 INTRODUCTION

Governments are interested in the development of policies, processes, standards in Information and Communication Technology (ICT), mounting structures dedicated to achieving interoperability. Mainly in digital preservation, the challenge of interoperability in addition to a technical is a social and institutional problem, as it depends on institutions that pass through changes of direction, mission, administration and funding sources. Therefore, the Brazilian government established a set of interoperability standards called e-PING (Padrões de Interoperabilidade de Governo Eletrônico)[1]. Concerning the way storing information, E-PING adopts document format Open Document Format(ODF) to transmit government information among public and private sectors maintaining privacy and security. Digital documents are considered official records and are managed according to laws and standards that understand entire lifecycle of these materials(Bretas and do Socorro Ferreira Mesquita, 2010).

Choosing a common standard format for exchanging information is excellent, but still continues dependent on a specific format, even being free. The main problem with digital documents is to ensure access to those documents in the long term. So it is necessary to overcome technical barriers associated with document formats. The content, structure and context of documents must be associated with software features that preserves its representations and relationships enabling their reconstruction. The purpose of this work is to facilitate the distribution of documents, overcoming the problem of formats with which they were created. Besides, we aim to enable documents interoperability and achieve documents portability simply through the reuse of content and formats in different plausible combinations.

Ontologies can assist us in this work. They provide a shared understanding of terms allowing interoperability and means for an intelligent integration of information(Uschold, 1998). This work uses ontologies as a solution to the lack of problem of interoperability in implementations of document formats. The proposal is to represent digital documents based on two ontologies: (1) format ontology, that characterizes the digital documents structure and presentation, independently of specific encoding of each software product and (2) context ontology, that represents the information contexts of businesses. Documents are offered based on generic representations centralized, through mediators among ontologies, presentation systems and document editing. As proof of concept will be considered interoperability between document formats ODF and OOXML.

[1] http://www.governoeletronico.gov.br/

The paper is organized as follows: in section 2 we explain the problem of preserving digital documents; in section 4 we introduce the fundamental concepts of Document Engineering; in section 3 we show the main concepts of document format. In section 5 we summarize some related work, and in section 6 we explain our proposed method. Then, in the section 7, we present the results. Finally, in section 8 we point out some closing remarks and future works.

2 PRESERVATION OF DIGITAL DOCUMENTS

Currently doing business by document exchange is natural and intuitive. Documents are interfaces for people and business processes. However, there is a lot of file are formats incompatible. Many documents written and stored with the same format can be unreadable, inoperable or after some time be necessary to migrate the data. Therefore organizations need to manage their knowledge effectively in order to preserve their intellectual capital. Organizations need to provide documents independently of the software created. Hence it, is important to enable documents interoperability within the business processes, i.e., to enable coherent exchange of information based on the adoption of rules and communication standards that allows comunication between heterogeneous systems.

When documents are exchanged, it is very important to ensure documents authenticity. Authenticity is the ability to identify diplomatic elements that allow assessing whether a document is authentic. A document is authentic if it can prove that there is a set of properties, considered significants, that were correctly preserved along time. To achieve authenticity is fundamental to rightly record the provenance of the document. Contextualize their existence, describe their custodial history and testify to their integrity was not compromised (Ferreira, 2006).

3 DOCUMENT FORMATS BASED XML

Document format encapsulates a complete description for storing digital documents. For instance, organizing elements such as text, fonts, graphics, and other information needed to display a document. Among the frequently used document formats are Office Open XML (OOXML) and Open Document Format (ODF). They are main open standards based XML for document formats. However, there are no implementations that offer 100% of portability for both, not even into the dominant implementations Microsoft Office and Open Office.org (Shah and Kesan, 2009). For example, to present some differences between them, Figure 1 shows a simple text encoded in standards OOXML and ODF.

3.1 Office Open XML (OOXML)

The specification OOXML exists in several versions. In 2008, a revised version of the ECMA-376 (2006) became an open standard ISO (ISO / IEC 29500:2008). OOXML has approximately 5000 pages and it defines formats for word-processing, presentation, and spreadsheet documents. Each type of document is specified through a primary markup language: WordprocessingML, PresentationML, or SpreadsheetML. Embedding mechanisms permit a document of any one of these three types to contain material in the other primary markup languages and in a number of supporting markup languages (Ngo, 2008).

3.2 Open Document Format (ODF)

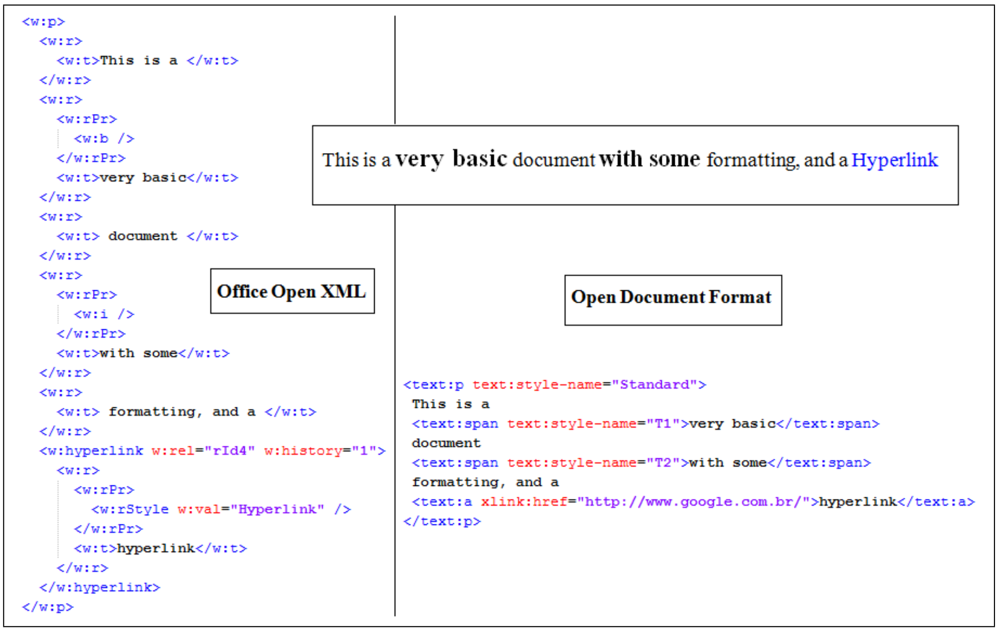
ODF or OD was originally developed by Sun Microsystems in 2000 for StarOffice and OpenOffice.org. ODF was designed with the goal of being a specification document format independent supplier or Software, making it suitable for storing documents long term.

ODF was designed for being a specification document format independent supplier or software, making it appropriate for storing documents long term. ODF was published in 2005 as OASIS standard[2](Organization for the Advancement of Structured Information Standards). In 2006, this standard was approved by the organizations as ISO/IEC 26300:2006 ODF v1.0 and in 2012 OASIS published version ODF v1.2. It has approximately 700 pages and it is used for documents containing text, spreadsheets, charts, and graphical elements. The file format makes transformations to other formats simply by leveraging and reusing existing standards wherever possible. ODF is the format more recommended by academy and defenders of open standards (Durusau et al., 2011).

4 DOCUMENT ENGINEERING

Documents are purposeful representations and organizations of information, but they exhibit great vari-

[2] http://www.oasis-open.org



```
<w:p>
  <w:r>
    <w:t>This is a </w:t>
  </w:r>
  <w:r>
    <w:rPr>
      <w:b />
    </w:rPr>
    <w:t>very basic</w:t>
  </w:r>
  <w:r>
    <w:t> document </w:t>
  </w:r>
  <w:r>
    <w:rPr>
      <w:i />
    </w:rPr>
    <w:t>with some</w:t>
  </w:r>
  <w:r>
    <w:t> formatting, and a </w:t>
  </w:r>
  <w:hyperlink w:rel="rId4" w:history="1">
    <w:r>
      <w:rPr>
        <w:rStyle w:val="Hyperlink" />
      </w:rPr>
      <w:t>hyperlink</w:t>
    </w:r>
  </w:hyperlink>
</w:p>
```

Office Open XML

This is a **very basic** document *with some* formatting, and a Hyperlink

Open Document Format

```
<text:p text:style-name="Standard">
 This is a
 <text:span text:style-name="T1">very basic</text:span>
 document
 <text:span text:style-name="T2">with some</text:span>
 formatting, and a
 <text:a xlink:href="http://www.google.com.br/">hyperlink</text:a>
</text:p>
```

Figure 1: Sample text encoded in formats OOXML and ODF

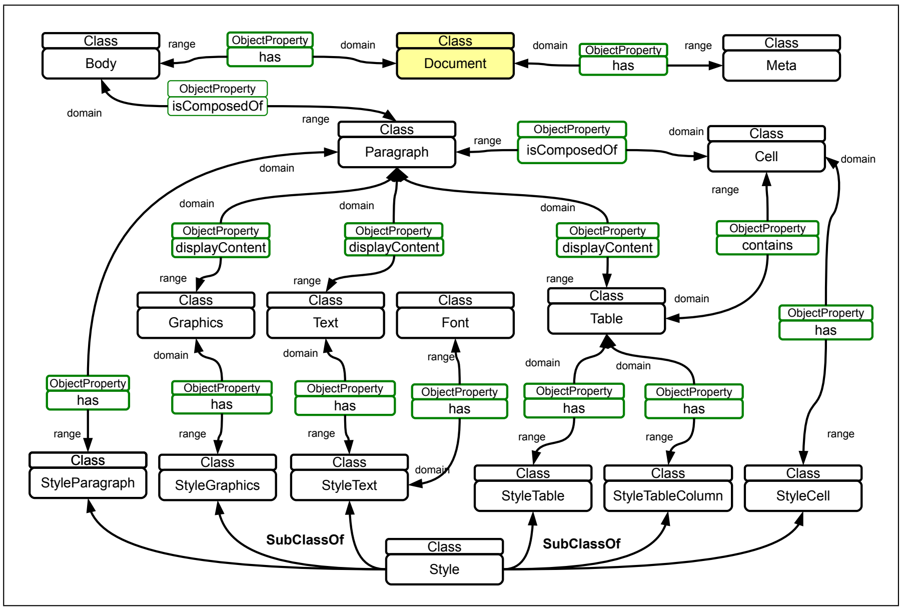


Figure 2: Format Ontology

ety. Document engineering analyse and design methods that yield precise specifications for the information and rules that business processes require. This mean, developing models that emphasize document requirements and patterns of information exchange, focusing the separation of content and presentation of the document information in an inherent and desirable way. A document must be represented using a shared common conceptual model (Glushko and McGrath, 2008).

Document engineering define models of different types of documents in a rigorous and unambiguous way so that we can automate their process or exchange within or between applications. It needs to be diligent and precise when defines the meaning of any information produced and consumed by business applications. As electronic documents are ubiquitous, document engineering emphasize that a document must be defined in a technology-neutral way as a purposeful and self-contained collection of infor-

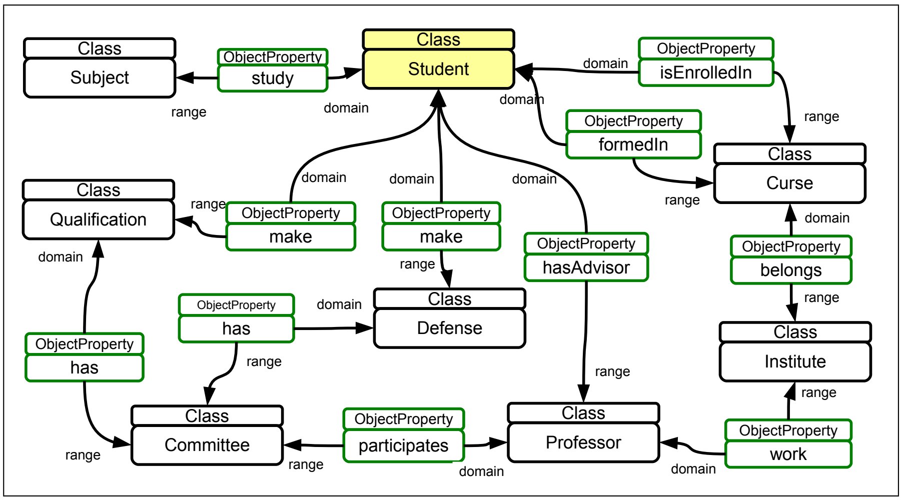


Figure 3: Context Ontology for Student Sheet

mation. When businesses exchange documents, they must agree on what the documents mean and on the business processes they expect each other to carry out with them, but they do not need to agree on the technology they use (Glushko and McGrath, 2008).

Many applications need to support different physical interfaces. These imply many-to-many mappings between the input and output interfaces for each application. Many-to-many mappings can be avoided by mapping all physical interfaces to a common conceptual model. So, the best way to facilitate interoperability is allowing the participants to share the same conceptual model. A common metamodel helps aligning different models. Basing the user and application interfaces on a common conceptual model ensures that the documents they process are interoperable (Glushko and McGrath, 2008).

Document engineering entails the need for standardization of syntax, structure, and semantics of business documents and their reusable components. Document engineering search achieve documents interoperability and improve documents portability. In other words, document interoperability is the ability of businesses applications to extract information contained in various kinds of documents and transform it standardized XML structures. These XML data files can then be exchanged between the various systems and further processed (Schmidt et al., 2006). Using standardized XML structures saves effort and yields more consistent, compatible, and successful designs. Moreover, document portability refers to the exchangeability of documents as a whole, i.e. with all the information they contain, formatting settings and graphic information. Crucially, all stylistic and graphical data (Schmidt et al., 2006). This information could be grouped in four components: content, logical structure, layout structure and presentation(Barron, 1996).

Document portability considers the issue of visual fidelity to an original, makes requirements in terms of optical appearance, stylistic elements and other such matters. On the other hand, document interoperability is exclusively concerned with the exchange of the business information contained in documents. Document interoperability could enable businesses applications to communicate directly with a wide range of different eGovernment services, platforms and administration applications. The document interoperability shall enable business processes to be generated from office applications, and then to be integrated in corresponding eGovernment processes(Schmidt et al., 2006).

5 RELATED WORKS

An excellent work existent in the literature is based on the framework UN/CEFACT CCTS[3](Core Component Technical Specification). CCTS is an european conceptual framework for modeling document components in a syntax neutral and technology independent manner. It permits handling different document configurations imposed by divergent national legislations. The conceptual model is transferred to XML schema serving as a basis for Collaborative Web Services for eGovernment(Vogel et al., 2008).

Another interesting work is designing XML documents from conceptual schemas and workload information for compliant to consensual information of

[3] http://www.unece.org/cefact/index.html

specific domains. The research presents a conversion approach which considers data and query workload estimated for XML applications, in order to generate an XML schema from a conceptual schema. Loaded information is used to produce XML schemas which can respond well to the main queries of an XML application. The work evaluates an approach through a case study carried out on a native XML database. Its experimental results demonstrate that the XML schemas generated by the proposed methodology contribute to a better query performance than related approaches(Schroeder and Mello, 2009).

Lastly, Universal Business Language (UBL) is a library of standard electronic XML business documents such as purchase orders and invoices. It was developed by OASIS. UBL is designed to provide a universally understood and recognized commercial syntax for legally binding business documents and to operate within a standard business framework such as ISO 15000 (ebXML) to provide a complete, standards-based infrastructure that can extend the benefits of existing EDI systems to businesses of all sizes. It is designed to eliminate the re-keying of data in existing fax- and paper-based business correspondence and provide an entry point into electronic commerce for small and medium-sized businesses. UBL schemas are modular, reusable, and extensible in XML-aware ways. UBL Library is based on a conceptual model of information components known as Business Information Entities (BIEs). These components are assembled into specific document models. One document is a set of information components that are interchanged as part of a business transaction; for example,in placing an order. These document assembly models are then transformed in accordance with UBL Naming and Design Rules into W3C XSD schema syntax. This approach facilitates the creation of UBL-based document types beyond those specified in this release(Bosak et al., 2011).

6 USING ONTOLOGIES FOR MODELING DIGITAL DOCUMENTS

Ontologies are designed for enabling knowledge sharing and reuse on some domain that can be communicated between people and computers. Therefore, to enable the sharing and reuse of knowledge, it is necessary a formal specification of concepts. Ontologies define rules of relationships between concepts to query, infer knowledge(Gruber, 1995).

The documents present information depending on its purpose, and this information can be presented in different ways. Our proposal is to build a model that considers essential qualities of digital document. Following the approach of engineering documents, a document is considered as combination of information components and presentation components. Whereas the information is independent of how it is presented. This model is based on a integration of two ontologies. An ontology that represents the presentation structure and an other that represents the information according business context, i.e. , a format ontology and a context ontology. These ontologies are independent between each other. The objetive of format ontology is to achieve simple document portability. The purpose of context ontology is to achieve document interoperability. Figure 4 shows our proposal. The mapping between the format ontology and content ontology for its physical interfaces occurs through translators.

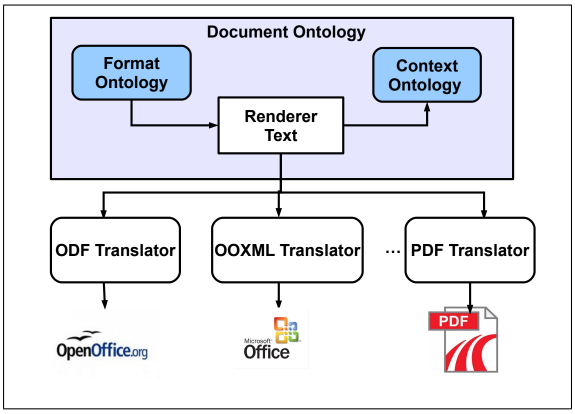


Figure 4: Document generation from format ontology and content ontology.

6.1 Format Ontology

Format ontology characterizes visual structure of document, i.e., formatting settings and graphic information. Format ontology specifies formally document components, i.e., presentation structure including metadata, paragraphs, texts, tables, lists, enumerations, images, styles, etc. The Metadata is the information associated with the document, for example: creation date, last modified date, text language, document author, pages number, etc. The document's layout is based on tables. The tables are composed of cells, cells can contain paragraphs with images, text ou maybe another table. Each paragraph of text has a presentation style, i.e., color, color-font, font, size, horizontal alignment, vertical alignment, etc. The format ontology is shown in Figure 2. From the format ontology, a document can be created in an appropriate

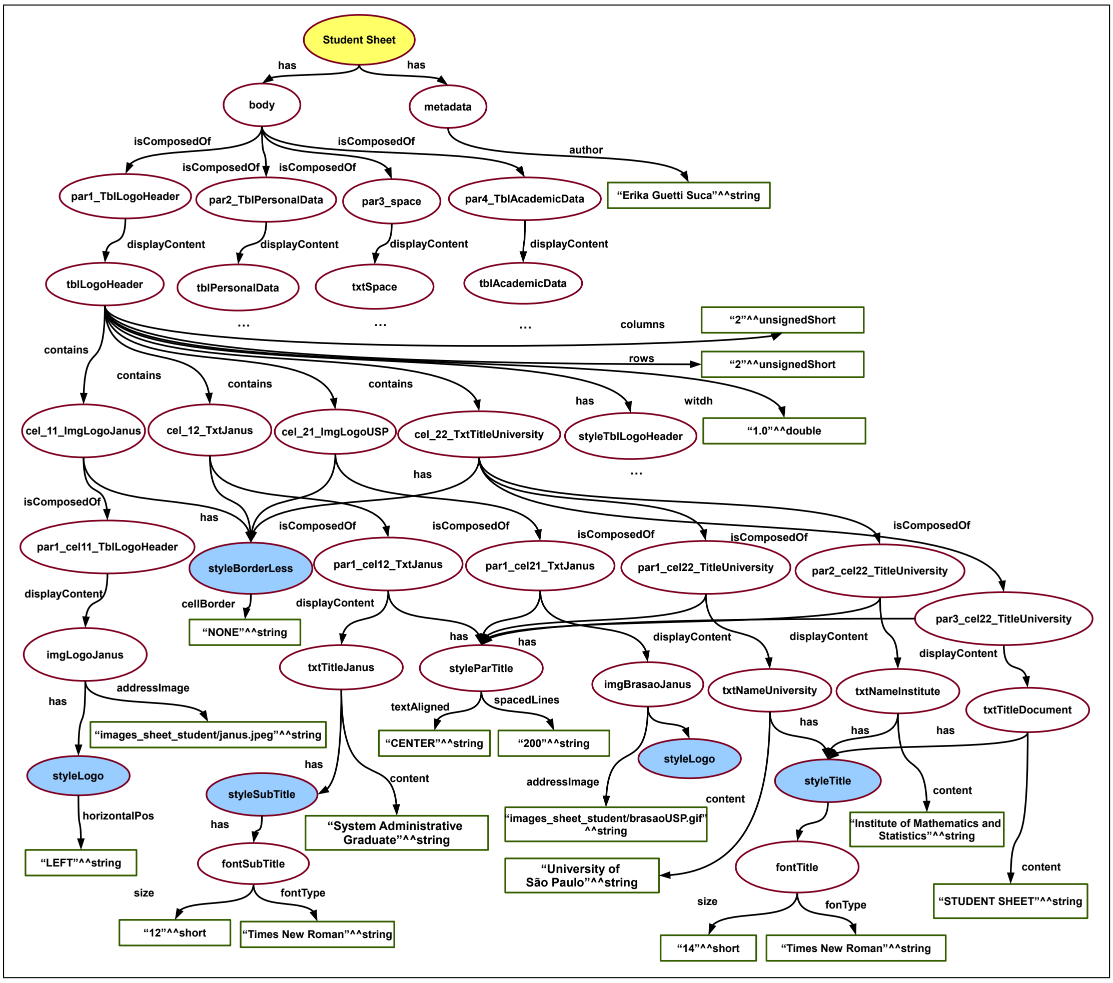
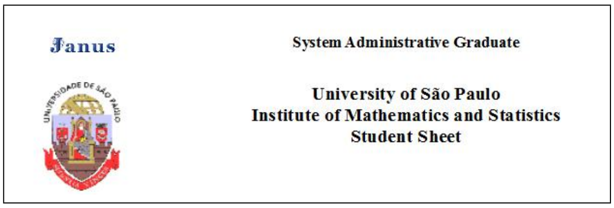


Figure 5: Tree instances of ontology format corresponding to header's student sheet



Janus

System Administrative Graduate

University of São Paulo
Institute of Mathematics and Statistics
Student Sheet

Figure 6: Sample of header's student sheet.

format for its purpose.

6.2 Context Ontology

A business context is a scope in which a specialized vocabulary is employed, so the business context defines the type of document information. The context is used to organize and analyze requirements and rules information presentation. For example, a student sheet, a medical record, rental contract, etc. show different scenarios. For proof of concept this work takes the context of a student sheet. The main objetive of a student sheet is to provide academic information, i.e., grades, attendance, subjects, advisor, date of birth, date of admission, etc. The context ontology was created based on that context. Context ontology is shown in Figure 3.

6.3 Generating Documents

Figure 4 summarizes the process of document generation from combination of instances of ontology format and content. However, the number of instances to represent a document usually is big. For example, Figure 5 shows a tree of instances of ontology format that characterizes the text shown in Figure 6.

A document is generated from the combination of format ontology and context. The instances of this ontologies are mapped into intermediate objetcs from which a document is created in an appropriate format. For example, Figure 7 shows an instance of concept Text and its value of attribute content. The string, *#institute_EN_#nameInstitution* contains the name of an instance and the name of its attribute separated by the symbol #. The concept Text rederizes in the document the value of attribute nameInstitution, that for this example is *University of São Paulo*. The instances of concept Text are always in a paragraph and a document is composed of paragraphs. So, the instances of ontology format refer to the instances of ontology content, which, in turn, are entries for mappers of ontologies to objects that represent components of docx and odf formats. Finally, based on the representations of docx and odf formats, a document is rendered in an appropriate format.

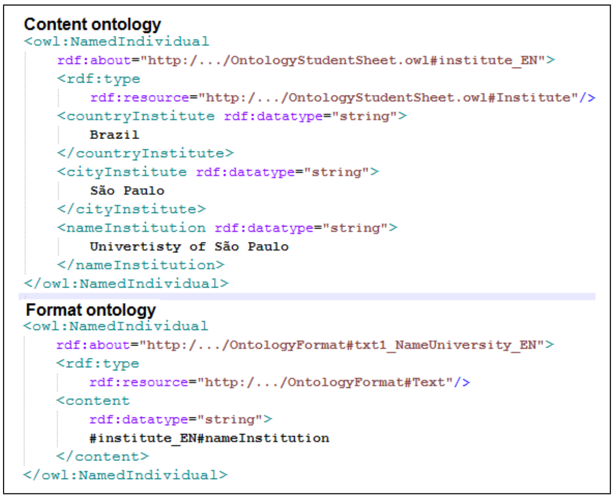
7 RESULTS

This study tested a small subset, basic word processing features, of what is needed for multiple interoperable implementations. This work is not trying to test extremely complex elements, but elements that are routinely used.

The experiment was implemented in java using OWL Api[4] 3.2, OpenDocument Odfdom [5] 0.8.7 and Apache POI[6] 3.8. The documents generated in ODF were tested using OpenOffice.org3. In the case of the format OOXML, the documents were tested using Microsoft Office 2010. Lastly, the ontologies were created using Protégé[7] 4.2.0.

The main difficulty in the experiment was maintaining the aesthetic characteristics fidelity of the document. We have not considered features not compatibles between standars OOXML and ODF. It is not possible to obtain always 100% of the translatability between OOXML documents to ODF and vice versa, due to the unique characteristics of standards OOXML and ODF(Eckert et al., 2009).

The objective was to enable sharing documents keeping the integrity of their information, i.e., to achieve document interoperability while allowing simple portability. The format ontology and content ontology overcome the problem of preserving digital documents, eliminating the dependence of particular technologies and enabling the central storage of documents. The context ontology can be reused in other applications and take full advantage of the power of expressivity of ontologies. The format ontology can generate documents regardless of context ontology. Figures 8, 9, 10 and 11 show the results of some experiments. Despite failing to reproduce the 100% of the stylistic features equally in both formats, the student's information integrity has not been compromised. The two student sheets continue serving their purpose, reporting the data and student performance. The documents maintained its authenticity.



Content ontology
```
<owl:NamedIndividual
    rdf:about="http:/.../OntologyStudentSheet.owl#institute_EN">
    <rdf:type
        rdf:resource="http:/.../OntologyStudentSheet.owl#Institute"/>
    <countryInstitute rdf:datatype="string">
        Brazil
    </countryInstitute>
    <cityInstitute rdf:datatype="string">
        São Paulo
    </cityInstitute>
    <nameInstitution rdf:datatype="string">
        Univertisty of São Paulo
    </nameInstitution>
</owl:NamedIndividual>
```

Format ontology
```
<owl:NamedIndividual
    rdf:about="http:/.../OntologyFormat#txt1_NameUniversity_EN">
    <rdf:type
        rdf:resource="http:/.../OntologyFormat#Text"/>
    <content
        rdf:datatype="string">
        #institute_EN#nameInstitution
    </content>
</owl:NamedIndividual>
```

Figure 7: OWL code of an instance of concept Text of format ontology and an instance of concept Institute of content ontology.

8 CONCLUSIONS

This work has shown that ontologies and simple mappings provide a good foundation for the creation of digital documents allowing document interoperability and simple portability. In addition, based on centralized representations of documents, it is possible to change the physical interface without changing all its mapped physical interfaces. Even more, new physical interfaces could be added, perhaps for new display or output devices, without changing the conceptual model. If technologies change, translators will also change even though the underlying conceptual models will not.

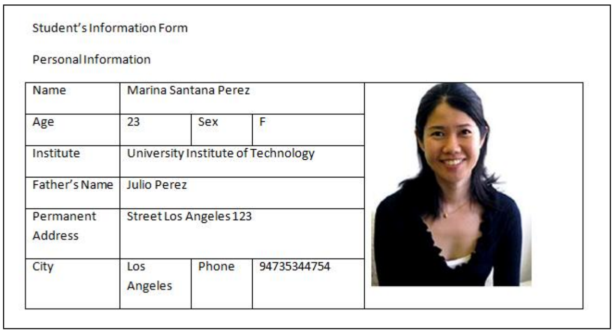
Using the same base of conceptual model, multiple publicationsand formats could be created, and different assemblies of document could share common

[4] http://owlapi.sourceforge.net/

[5] http://incubator.apache.org/odftoolkit/odfdom/index.html

[6] http://poi.apache.org/

[7] http://protege.stanford.edu/

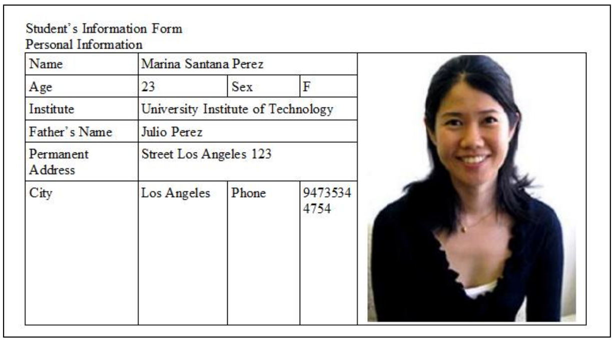


Student's Information Form

Personal Information

Name	Marina Santana Perez		
Age	23	Sex	F
Institute	University Institute of Technology		
Father's Name	Julio Perez		
Permanent Address	Street Los Angeles 123		
City	Los Angeles	Phone	94735344754

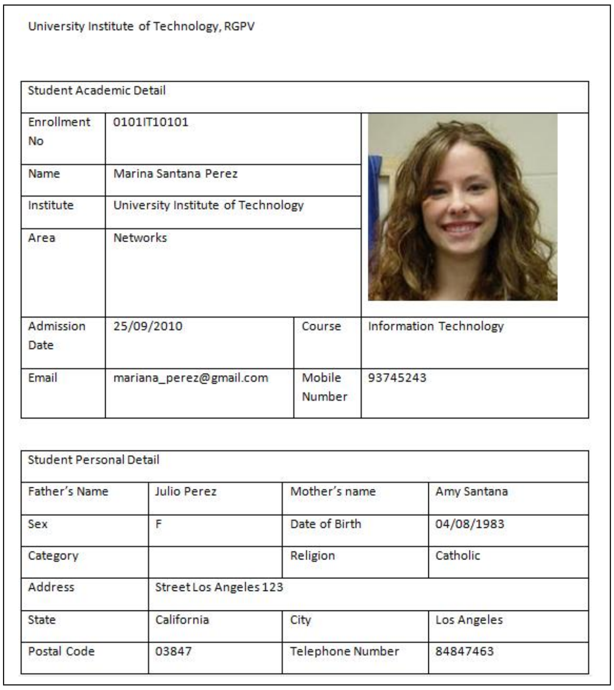
Figure 8: Sample 1, version generated for DOCX format.



Student's Information Form
Personal Information

Name	Marina Santana Perez		
Age	23	Sex	F
Institute	University Institute of Technology		
Father's Name	Julio Perez		
Permanent Address	Street Los Angeles 123		
City	Los Angeles	Phone	94735344754

Figure 9: Sample 1, version generated for ODT format.

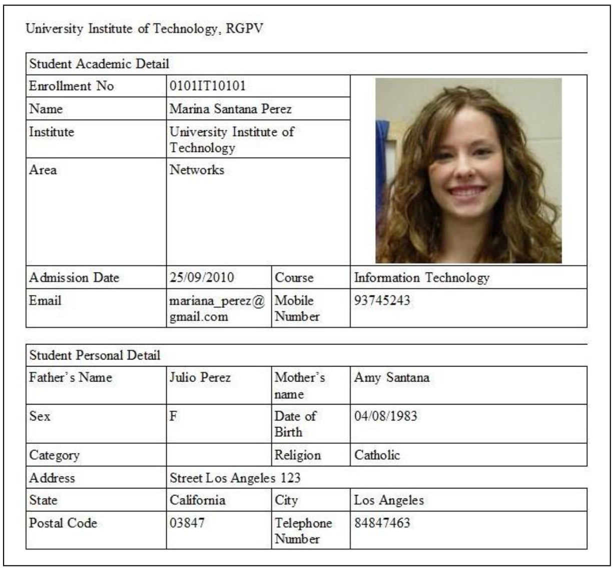


University Institute of Technology, RGPV

Student Academic Detail			
Enrollment No	0101IT10101		
Name	Marina Santana Perez		
Institute	University Institute of Technology		
Area	Networks		
Admission Date	25/09/2010	Course	Information Technology
Email	mariana_perez@gmail.com	Mobile Number	93745243

Student Personal Detail			
Father's Name	Julio Perez	Mother's name	Amy Santana
Sex	F	Date of Birth	04/08/1983
Category		Religion	Catholic
Address	Street Los Angeles 123		
State	California	City	Los Angeles
Postal Code	03847	Telephone Number	84847463

Figure 10: Sample 2, version generated for DOCX format.



University Institute of Technology, RGPV

Student Academic Detail			
Enrollment No	0101IT10101		
Name	Marina Santana Perez		
Institute	University Institute of Technology		
Area	Networks		
Admission Date	25/09/2010	Course	Information Technology
Email	mariana_perez@gmail.com	Mobile Number	93745243

Student Personal Detail			
Father's Name	Julio Perez	Mother's name	Amy Santana
Sex	F	Date of Birth	04/08/1983
Category		Religion	Catholic
Address	Street Los Angeles 123		
State	California	City	Los Angeles
Postal Code	03847	Telephone Number	84847463

Figure 11: Sample 2, version generated for ODT format.

structures or patterns. The idea is to reuse standard schema components wherever it is possible. We could easily imagine applications that reuse documents and processes in ways not anticipated by theirs creators.

Part of future work is to improve the implementation of the translators and include others formats, PDF, HTML, etc. We will also develop a use case document interoperability applied to electronic government, where the format ontology and content ontology should play an important role in preserving and distributing digital documents efficiently. The creation of instances of ontologies is a long and laborious process, so, another future work would be the consolidation of a tool that assists in automating the creation of instances.

REFERENCES

Barron, D. W. (1996). Portable documents: problems and partial solutions. *Department of Electronics and Computer Science University of Southampton*, 8:343–367.

Bosak, J., McGrath, T., and Holman, G. K. (2011). Universal business language v2.1. Organization for the Advancement of Structured Information Standards (OASIS), Standard.

Bretas, N. L. and do Socorro Ferreira Mesquita, C. (2010). *Panorana da Interoperabilidade no Brasil*. Ministério do Planejamento, Orçamento e Gestão.

Durusau, P., Brauer, M., and Corporation, O. (2011). Open document format for office applications v1.2. Organization for the Advancement of Structured Information Standards (OASIS), Standard.

Eckert, K.-P., Ziesing, J., and Ishionwu, U. (2009). *Document Operabilitiy Open Document Format and Office Open XML*. Fraunhofer Verlag, Germany, fokus edition.

Ferreira, M. (2006). *Introdução e preservasão digital: Conceitos, estratégias e actuais consensos*. Escola de

Engenharia da Universidade do Minho, edição electrônica edition.

Glushko, R. J. and McGrath, T. (2008). *Document Engineering: Analyzing and Designing, Documents for Business, Informatics and Web Services*. Massachusetts Institute of Technology.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, 43(5-6):907–928.

Ngo, T. (2008). Visão geral do Office Open XML. Technical report, ECMA TC45.

Schmidt, K.-U., Fox, O., Henckel, L., Holzmann-Kaiser, U., Martin, P., and Tschichholz, M. (2006). *Document Interoperability for Use in eGovernment: Integration of XML-based Document Content in Public Administration Processes*. FOKUS.

Schroeder, R. and Mello, R. D. S. (2009). Designing xml documents from conceptual schemas and workload information. *Multimedia Tools Appl.*, 43(3):303–326.

Shah, R. and Kesan, J. (2009). Interoperability challenge for open standars: Odf and ooxml as examples. *The proceedings of the 10th International Digital Government Research Conference*.

Uschold, M. (1998). Knowledge level modelling: concepts and terminology. *The Knowledge Engineering Review*, 13:1:5–29. Printed in the United Kingdom.

Vogel, T., Schmidt, A., Lemm, A., and Österle, H. (2008). Service and document based interoperability for european ecustoms solutions. *J. Theor. Appl. Electron. Commer. Res.*, 3(3):17–37.