E-MEDISYS 2008
2nd International Conference: **E-M**edical **S**ystems
October 29-31, 2008 – TUNISIA

IPSH: Semantic Integration of Biomedical Analysis Data Deriving from Image Processing

André Bouchot*, Christophe Cruz**, Franck Marzani**, Christophe Nicolle**

* *Plateforme d'Imagerie Cellulaire, IFR 100, Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France*
bouchot@u-bourgogne.fr

** *LE2I UMR CNRS 5158, IFR 100, Université de Bourgogne, BP 47870 21078 Dijon Cedex, France*
{christophe.cruz, fmarzani, cnicolle}@u-bourgogne.fr

Abstract: Histology services, electronic microscopy services and optical microscopy services use heterogeneous and sophisticated tools to analyze biomedical samples. A huge quantity of data, images and videos are produced during these analyses. The project IPSH aims at semantically integrating data from biomedical analyses and image processing. Image processing is undertaken in order to determine the division rate in these cells and to verify the implication of the transfected protein. Each identified cell in one image is linked to the other images of the same sequence and is also linked to the knowledge base in order to access advanced data.
Key words: Biomedical analysis, image processing, data integration, knowledge management, ontology.

INTRODUCTION

Histology services, electronic microscopy services and optical microscopy services use heterogeneous and sophisticated tools to analyze biomedical samples. A huge quantity of data, images and videos are produced during these analyses. The important process during this production consists in localizing, identifying and characterizing anomalies. During the last decade, new image processing algorithms have been developed to reduce noise for a better representation of biological elements. Most of them make it possible to identify biological elements on pictures which can be, for instance, a group of cells. Due to the heterogeneity of data, there is, today, no process that allows patricians to connect these elements to various kinds of data or processes in order to build a coherent knowledge database.

In addition, the development of networks increases the need of software packages and workflow methods in order to permit the collaboration of specialists. Unfortunately, data heterogeneity and misinterpretation of information limits the systems' interoperability. Our goal is to build a semantic framework which supports the sharing and the combination of heterogeneous data. Concisely, the IPSH project aims at semantically integrating data from biomedical analyses and image processing. In order to support this integration, we are currently developing a platform that is based on knowledge management technologies and that permits to index and classify data from histology services, electronic microscopy services and optical microscopy services. Data generated in these services are sheets, series of pictures according to time, wavelength, and depth. In order to support the image processing, for instance, from a Hela cell culture that has been transfected for a particular protein, we use an automatic process that aggregates the neighboring pixels of cells. These aggregates are realized from a sequence of "phase contrast" images. This process is undertaken in order to determine the division rate in these cells and to verify the implication of the transfected protein This process cannot be easily carried out manually due to the fact that more than one thousand images are generated during the data acquisition phase. In addition, each identified cell in one image is linked to the other images of the same sequence and is linked to the knowledge base in order to access advanced data.

This paper focuses on the data integration and on knowledge management. There are no details on the automatic process of knowledge extraction from images.

1. Background

Over the past several years many companies have turned to commercial ETL (Extract, Transform, and Load) tools as a means to reduce the effort associated

with the most common integration approach: manual coding. Using a centralized ETL "engine" as an integration hub, and powered by a single, platform-independent language, most of the ETL tools reduce the effort associated with point-to-point manual integration. The ETL solutions are based on research undertaken since the 80s on distributed database systems [BAT 84]. Unfortunately, this solution is limited to the syntactic and schematic data descriptions in integration processes. Various computer science areas, from the artificial intelligence to the most recent Semantic Web, while passing by software engineering, propose solutions to represent knowledge in order to make the data comprehensible, easy to handle and manageable by an information processing system. The ultimate solution to solve the problems concerning data interoperability is to impose a standard. This has been done during the development of the Web with protocols such as TCP/IP, HTTP, FTP and HTML standard. With the emergence of Information Technologies (IT), all actors of the Web have defined a common structure for data which is called XML.

This technology has emerged as an important model to describe and share Web based data and processes. Its importance is provided by two major factors. First, XML became de facto a data standard supported by many software vendors and application developers. Second, XML is based on a relatively simple structure which is readable by both, user and machine. In addition, it can be used by many non-professional users who are not expert in database administration. During the last five years, XML has known an incredible and an extensive use by systems that exchange and share data. XML is now used to structure most of the Web information (videos, images, 3D scenes, domain information, office data, etc.) Many systems using XML, such as database integration systems, have a mediation approach [CAL 01], [CAR 00], [RAP 02]. The evolution of the Web technologies has changed the problem of information integration. Indeed, the contribution of XML to the definition not only integrated schemas but also of languages for the corresponding models has considerably reduced the problems related to the syntactic and the schematic heterogeneity. Nevertheless, during the integration data process and the integration services there remain many problems related to semantic heterogeneity. Consequently, a formal description of the shared semantic should avoid ambiguities when it is defined in relation to a domain of knowledge. Hence, the reuse of information in a specified domain ought to be improved and ought to facilitate the knowledge management.

Knowledge of documents has traditionally been managed through the use of metadata. The Web semantic proposes to annotate the document content using semantic information from domain ontologies [RDF 04]. The result is a set of Web pages interpretable by a machine with the help of mark-ups. The goal is to create annotations (manually or automatically) with well-defined semantics. In the Semantic Web context, the content of a document can be described and annotated using RDF and OWL. Resource Description Framework (RDF) [BER 01] is a formalism of knowledge representation from the semantic networks field. It is mainly used to describe resources, such as web documents, with a set of metadata (author, data, source, etc.) and a set of descriptors. This metadata is composed of a triplet: (object1, relationship, object2) or (resource, property, value). Web Ontology Language (OWL) [OWL 08] is used to specify ontology or, more generally, some ontological and terminological resources by defining concepts used to represent a domain of knowledge. Each concept is described by a set of properties, relations and constraints. The OWL formalism is derived from the Description Logic fields and has the capability to infer new knowledge from existing knowledge.

Semantic Web annotation brings benefits of two kinds to this platform - enhanced information retrieval and improved interoperability. Information retrieval is improved by the ability to perform searches which exploit the ontology in order to make inferences about data from heterogeneous resources [WELL 99].

Semantic Web standards for annotation tend to assume that the documents which are to be annotated are in Web-native formats such as HTML and XML. Annotea [KAH 01] is a W3C project which specifies infrastructure for annotation of Web documents. The Annotea framework was used in various tools including Vannotea [SCH 03]. The CREAM framework [HAN 02] specifies components required by an annotation system including the annotation interface. These approaches will have limited usefulness for knowledge management in this project. Actually, documents will be in many formats which are not XML based [URE 06]. The project LabelMe is a Web-based annotation tool for images that provides a drawing interface [RUS 08]. Based on the ontology called Wordnet[1], the users label images by clicking on them and by adding a key word. The users are free to label as many objects depicted in the image as they wish. The knowledge managed by this framework is a terminological definition of graphical objects. Consequently, it is not possible to define an object or an instance of a class which can be found in several documents. Our platform takes into account this problem by defining an object for each cell which can be found in several images.

Our main objective is to build a collaborative platform related to the processes of cellular vision. Actually, these processes are not limited to the video acquisition and image processing aspects but include the traceability, storage, dissemination and enrichment of knowledge. In order to reach this goal, data are modelised in XML which permits the necessary syntactic interoperability. In addition, this platform is based on semantic-based indexation and annotation methods of XML documents. This method makes it possible to build a knowledge base for the field of the

[1] Wordnet: http://wordnet.princeton.edu/

cellular vision. This method makes it possible to integrate heterogeneous data (business skill, measures, 2D graphic, etc.) produced by the various actors intervening in these processes.

2. Data integration

XML data is composed of XML documents and XML schemas which validate the corresponding XML documents. Actually, an XML schema contains the knowledge that the application has to share through data exchange. However, XML covers only the syntactic level but does not support the semantic level. The semantic level describes the meanings between the input and output. The syntactic level is a set of rules that allows to create a sentence which will give the computer a set of instructions in order to complete a particular task. The lexical level deals with input device dependencies in which the user will specify the exact syntax [THO 02]. In the context of computer and information sciences an ontology defines a set of representational primitives which allow to model a domain of knowledge or discourse [GRU 08]. Ontologies are widely used to capture and organize knowledge concerning a particular domain. According to the latest version of the Semantic Web stack, it seems that RDF and OWL are languages adopted to define ontologies on the Web.

The knowledge defined in ontologies is used as an index to retrieve specific data [GAR 05], to infer new knowledge [SWR 04], to semantically annotate multimedia data [CAS 07], to find out Web Services automatically [MAR 07], or to match knowledge with other knowledge for a more general purpose. XML schemas contain the knowledge of a domain that was specified by the author. This specification is only syntactic without any semantical definition. This is due to the fact, that XML data are used to exchange data between processes that were developed for this data. In order to permit the exploitation of the knowledge contained in XML schemas and instances, we propose an ontology enrichment and an automatic population process from XML data based on a manual mapping of XML schemas. Ontology enrichment is the activity of extending an ontology by adding new elements (e.g. concepts, relations, properties, axioms) [CAS 07]. The enrichment process in our solution consists in annotating knowledge which is contained in XML schemas in order to define the ontology schema [FAA 04]. The knowledge extract here is manual and is carried out by the user who is the only one that knows which part of the XML schema will be required in future processes. The ontology population process is the activity of adding new instances to an ontology [CAS 07]. This process can be carried out automatically by an automatic process.

According to [CRU 04a], [KLE 02], [LAK 03], [CRU 06], data integration can be undertaken by defining rules of mapping between information sources and the ontological level. These rules consist in adding a semantic layer to source elements and, thus, provide a semantic definition to elements with regard to a consensual definition of the meaning. For that purpose, ontologies are useful in order to define a common semantic. Furthermore, schema matching is a well studied field that allows to automatically find out identical resources in the different schemas. Schema matching is a manipulation process on schemas that takes two heterogeneous schemas as input and produces as output a set of mapping that identify relations between the elements of the two schemas [THA 08]. This is required in many database applications, such as integration of web data sources, data warehouse loading and XML message mapping. Actually, we did not focus on the schema matching between ontology generated from several XML schemas. Our process consists in mapping XML schema elements to an existing ontology in order to enrich the ontology and to automatically populate it. Then, the populated ontology can be used to find out new knowledge, to extract a new XML schema with its associated XML document.

The next section gives an overview of the environment that concerns the project IPSH. Section 4.1 presents our method based on the field of ontologies and formal languages. Section 4.2 introduces the notion of semantic marks which is the base of our method.

3. Environment

This section is made up of two parts. The first part presents the activities of the technical platform and its distribution in three services. These services generate a set of heterogeneous data. These data must be coupled with the graphic results obtained using image processing algorithms presented in the second part.

3.1. Services of the technical platform

The technical platform is structured in three services:

1. The Service of histology which prepares the biological samples for an analysis in optical microscopy (white light or fluorescence). The reception of the samples is accompanied by an information sheet for each sample; i.e. 20 tumors, for example, represent 20 different information sheets. These sheets contain various information on the sample owner and the protocol of treatment (fixation type, inclusion type, cutting, coloring, or immuno-labeling). These data will be converted into XML data in order to allow interoperability between the processes.
2. The Service of electronic microscopy where the biological samples are prepared and where the analysis of these samples is carried out. The same system of sheets was set up for this service. The pictures which were taken at this point of the observation are indexed, recorded and are sent to the customer.
3. The Service of optical microscopy and analysis of images consists of the Cell Observer station from Zeiss. This station is an inverted microscope which has two sources of excitation (HBO lamp

and Polychrome V) and two filter wheels (one in excitation and one in emission). It is motorized and equipped with a numerical camera (HRm, Zeiss). This station makes it possible to automatically acquire series of pictures according to time, wavelength, and depth in the sample. This software package permits also to make analyses or calculations of a picture (2D or 3D space deconvolution, colocalization analyses, arithmetic on images, analyzes on grey levels). A system of classification and information has been adopted. For each document generated in these services we build the corresponding XML schema. These schemas will be used in our indexation process to build the final ontology.

3.2. Image Processing to SVG

The image acquisition consists in the capture of a sequence of images. Each of these images describe the biological sample according to the time. We have shown, for example, in figure 4, a Hela cell culture that has been transfected for a particular protein. We have acquired a sequence of phase contrast images in order to determine the division rate in these cells and to verify the implication of the transfected protein. No filter is used for this kind of application. In order to validate this project, some very classical methods of image processing have been implemented. Two kinds of segmentation are proposed.

The first one is manual-based. The user of the platform draws the borders of interesting cells in the sequence of images. The result is a set of pixels for each cell from each image. The identified zones are then modeled into vectorial data. The second segmentation method is semi-automatic. The user points out one pixel as a seed of an interesting cell. Then, an automatic process aggregates the neighboring pixels. The result is an area which defines the complete cell. Again, this zone is modeled into vectorial data. This process is repeated for all the cells to be discriminated in the image sequence.

Whatever the sample and the conditions of acquisition (filters or not), some image processes can be used. The result is a set of areas of interest. Each one is described by a set of pixels with grey or color values. These identified zones can be finally modeled into vectorial data according to an XML schema an SVG typed. At this stage, data which are available are histology XML documents and SVG documents pictures. In order to identify the same element in each document, some semantic links have to be defined. These links have also to refer to a common semantic definition which will be defined in an ontology of domain. The method used for creating links between documents is described in the following section.

4. Method overview

The principle of our solution consists in annotating and linking various levels which are the semantic level (OWL schema) and the schematic level (XML schema). The method is articulated in two steps.

The first step relates to the formalization of "mapping" rules between an XML schema and an OWL XML schema. The "mapping" rules make it possible to enrich an existing ontology of domain from concepts and relationships conceptually present in XML schemas. This can be realized by a machine. Semantically it will not be richer than the XML schema. Actually, a rich semantic mapping cannot be done by a machine for the moment. In addition, an ontology makes it possible to link the concepts and the relations from several schemas by amalgamating the attributes of common entities with the help of an identical semantic (figure 1 (E)). This step relates also to the definition of "advanced" rules in order to translate complex data such as sub trees into simple or complex attributes of a class.

The second step consists in populating the ontology previously enriched from XML documents and validated by the mapped XML schema (figure 1 (P)). The population has to follow some rules, such as, the limitation of cardinalitie attributes and unique instances. Consequently, "advanced" rules have to model and specify restriction on attributes. This principle will be described in the next section.

In order to specify the relevant elements of an XML schema for the enrichment process, it is necessary to identify and mark these elements. These marks are called "schematic marks" and are external RDF annotations of XML structures. (RDF rules in figure 1). We do not use the term "annotation" to designate a mark because we make the distinction between internal annotations that are added to the document and external annotations that are stored in a repository outside the annotated document. Actually, the "schematic marks" are a specific RDF graph that defines external annotations on XML schemas.

Once all the data is available in the knowledge base, information is available through the use of a request language. This language can be SPARQL[2] or any other language that defines a request for RDF data.

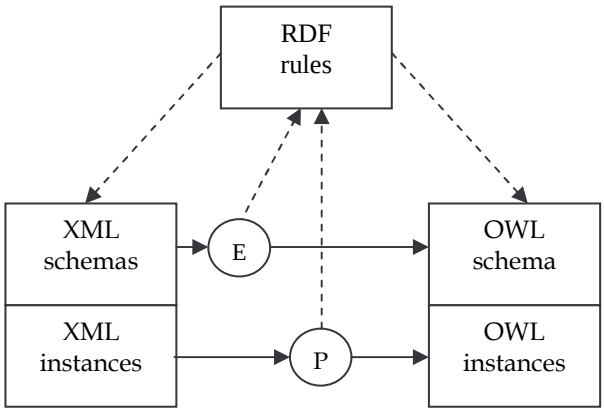

Figure 1. *Enrichment and population of an existing ontology. (E) is the process of enriching an ontology and (P) is the process of populating the ontology. These two processes use an RDF graph as rules to enrich and to populate the ontology. The rules in RDF are defined during the mapping process.*

[2] SPARQL, http://www.w3.org/TR/rdf-sparql-query/

The following section presents a brief formalization of schematic marks.

4.1. Factors and Schematic Marks

The properties of Dyck's languages were the subject of studies undertaken by J. Berstel [BER 00]. By drawing parallels between XML grammars and the languages of Dyck, J. Berstel defines the concept of "factor". This notion means that a language is factorizable in an under language and a factor of a Dyck's language is a Dyck's language. As a consequence, an under tree of an XML document can be generated by a factor of a Dyck's language to which the XML document belongs. According to the corollary 3.4, the notion of surface is used by Berstel to demonstrate the following proposition:

Proposition: For each XML language L there is only one reduced XML grammar generating L.

This proposal implies that if a factor were defined on an XML language then this factor would correspond to a production rule of the reduced XML grammar that generates this language. A reduced grammar does not have any useless non terminal vocabulary. An XML schema does not contain unnecessary tags, so an XML schema does not use unnecessary non terminal vocabulary. Consequently, an XML grammar is necessarily its own reduced grammar. This proposal makes it possible to introduce the concept of schematic mark.

Definition: A schematic mark is a mark on an XML schema that identifies a production rule. Each tag of an XML instance which has the same name was produced by the same production rules.

These marks are used in RDF rules to identify the production rules of the XML schema and are specified in XPath. Below, a non exhaustive list of production rules is given which can be marked and linked to an RDF mapping rule.

List of production rules:

xsd:schema	xsd:sequence	xsd:restriction
xsd:element	xsd:all	xsd:extension
xsd:attribute	xsd:group	xsd:minOccurs
xsd:attributGroup	xsd:choice	xsd:maxOccurs
xsd:complexType	xsd:any	xsd:minInclusive
xsd:simpleType	xsd:anyAttribute	xsd:maxInclusive
xsd:complexContent	xsd:union	xsd:minLength
xsd:annotation	…	xsd:maxLength
…		xsd:enumeration
		xsd:pattern
		…

4.2. Semantic of the schematic marks

An OWL grammar is also an XML grammar which can be marked as an XML schema. An ambiguous point has to be underlined. The OWL schema concerns the schema part of the ontology. However, OWL has an XML schema in order to validate OWL documents [URE 99]. In this section we focus on the XML schema of the OWL XML schema. This allows us to mark an OWL ontology, such as an XML schema. Consequently, we are able to mark production rules of the OWL ontology. Below, a non exhaustive list of production rules is given which can be marked and linked to an RDF mapping rule.

List of elements which are marked:

owl:Ontology	owl:disjointWith
owl:Class	rdf:type
rdfs:subClassOf	dfs:comment
owl:Objectproperty	owl:AnnotationProperty
owl:DatatypeProperty	rdfs:label
rdf:resource	owl:restriction
rdfs:domain	owl:onProperty
rdfs:range	owl:minCardinality
owl:intersectionOf	owl:maxCardinality
owl:unionOf	owl:inverseOf
owl:complementOf	owl:FunctionalProperty
owl:oneOf	owl:SymmetricProperty
owl:sameAs	owl:TransitiveProperty
owl:differentFrom	owl:hasValue
…	…

OWL uses most of the built-in XML schema data types. References to these data types are by means of the URI reference for the data type, http://www.w3.org/2001/XMLSchema. The following data types are recommended for the use with OWL:

List of data types:

xsd:string	xsd:int	xsd:gYearMonth
xsd:normalizedString	xsd:short	xsd:gYear
xsd:Boolean	xsd:byte	xsd:gMonthDay
xsd:decimal	xsd:unsignedLong	xsd:gDay
xsd:float	xsd:unsignedInt	xsd:gMonth
xsd:double	xsd:unsignedShort	xsd:anyURI
xsd:integer	xsd:unsignedByte	xsd:token
xsd:nonNegativeInteger	xsd:hexBinary	xsd:language
xsd:positiveInteger	xsd:base64Binary	xsd:NMTOKEN
xsd:nonPositiveInteger	xsd:dateTime	xsd:Name
xsd:negativeInteger	xsd:time	xsd:NCName
xsd:long	xsd:date	

4.3. Example of the schematic marks

Figure 2 shows an example of an XML schema that defines the kind of data generated by an analysis process made on a sample. This XML document defines the syntactic element which is called "Analysis". By marking it with an RDF triplet it is possible to specify the class "Analysis" in the ontology. In addition, all instances of this schema that contain an "Analysis" will be identified as an instance of the class "Analysis".

The TriG Syntax [TRI 08] is used to facilitate the explanations of how we employ schematic marks with XML schemas. Example 1 shows how to mark a first schema and how to specify an id in order to avoid duplicate OWL instances. The "bmr" name space is used to identify the basic mapping rules. In addition, "bmr:xpath" defines an element in the XML schema as does "bmr:xpathId" with the addition that the latter acts as an identifiant. A unique property on an XML element is welcome to define an identifiant.

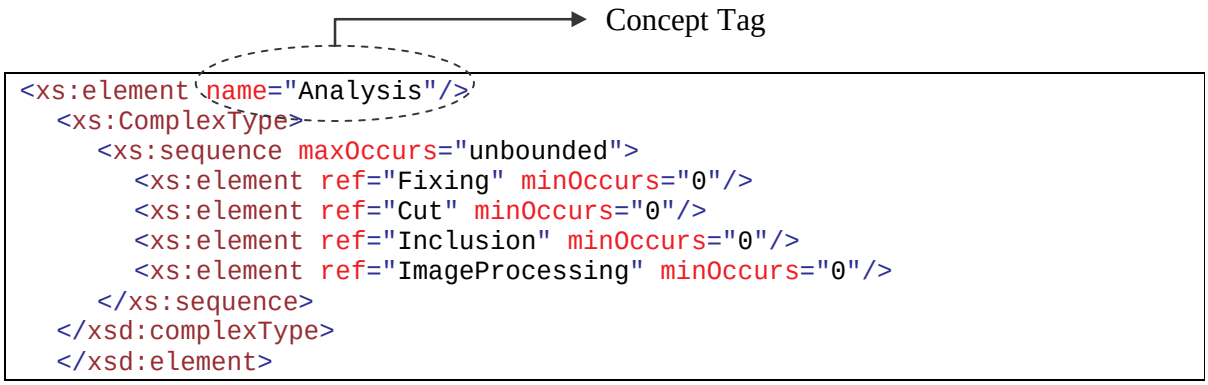

Figure 2. *In this figure the definition of a schematic mark for an Analysis tag is underlined. This tag is defined as a concept in the ontology. This mark defines a relationship between the XML document and the ontology of domain for data integration.*

```
# TriG Document 1
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix bmr: <http://www.example.org/BasicMappingRules#> .
:XSDMarksSchema_Cell {
:xml_cell       bmr:xpath          "/IM_PROC/CELL" .
:xml_cell       bmr:xpathId        "/IM_PROC/CELL[@IndexCELL]" .
}
```

Example 1. *This is an example of schematic mark definitions on an XML schema. It represents a document that contains information on cells.*

Example 2 shows how to mark a second schema. The Id element is specified.

```
# TriG Document 2
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix bmr: <http://www.example.org/BasicMappingRules#> .
:XSDMarksSchema_Analysis {
:xml_analysis    bmr:xpath   "/imageSequence/analysis " .
:xml_ana_cell  bmr:xpathId  "/imageSequence/analysis[@analysisId]".
}
```

Example 2. *This is an example of schematic mark definitions that represent a document about the defined analysis.*

Example 3 shows how to mark the OWL ontology. The elements "bmr:domain" and "bmr:range" permit to define relationships between predicates of triplets.

```
# TriG Document 3
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2001/XMLSchema#> .
@prefix bmr: <http://www.example.org/BasicMappingRules#> .
:OWLMarks_Cell {
:cell          bmr:xpath      "/rdf:RDF/owl:Class[rdf:ID='Cell']" .
:cellId        bmr:xpathId  "/rdf:RDF/owl:datatypeProperty
                                        [rdf:ID='indexCell']".

:analysis      bmr:xpath      "/rdf:RDF/owl:Class[rdf:ID='analysis']".
:analysisId  bmr:xpathId   "/rdf:RDF/owl:datatypeProperty
                                        [rdf:ID='analysisId']".

:hasCell       bmr:xpath     "/rdf:RDF/owl:objectProperty
                                        [rdf:ID='hasCell'].
:hasCell       bmr:domain          :analysis
:hasCell       bmr:range           :cell
}
```

Example 3. *This is an example of schematic mark definitions that represents the knowledge specified with an OWL ontology. It gives information on cells and analysis. "bmr:xpathId" is a type that defines an attribute which is an id.*

Conversion rules can also be also defined and consist in defining rules in order to convert properties that are different from the type of the property in the ontology and that cannot be directly copied into the ontology. The first kind of conversion is simple because it is the conversion of a simple type into another simple type (e.g. example 4).

```
:conversionBasicRules_1 {
:string2int     bmr:conv          :rule1 .
:rule1          bmr:type          bmr:convSimple .
:rule1          bmr:domain        xsd:string .
:rule1          bmr:range         xsd:int .
}
```

Example 4. *This is an example of a simple conversion from a string into an integer. The rule expresses only the semantic. The conversion process has to be done by a program. A set of conversion methods has to be carried out once and for all and can be reused for complex conversion rules.*

In order to define the semantic of schematic marks, these marks have to be linked to a semantic definition. It consists in defining relationships between XML schematic marks and OWL schematic marks. The mapping graph is also defined in RDF (e.g. example 5). The name space "amr" represents the "advanced mapping rules" elements.

```
# TriG RDF mapping rules
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix bmr: <http://www.example.org/BasicMappingRules#> .
@prefix amr: <http://www.example.org/AdvancedMappingRules#>.
mappingRules_1 {
:mapRule01    amr:domain        :xml_cell, xml_ana_cell .
:mapRule01    amr:range         :cell.

:mapRule02    amr:domain        :xml_analysis .
:mapRule02    amr:range         :analysis .

:mapRule03    amr:range         :cellId .
:mapRule03    amr:domain        :xml_cellId .
:mapRule03    amr:basicRule     :string2int.
}
```

Example 5. *In this example, "xml_cell" and "xml_ana_cell" are mapped to the class cell with the*

attribute cellId (with a conversion rule :string2int). In addition, "xm_analysis" is mapped to the class "analysis". Concerning the relationships between "analysis" and "cell", the ontology contains already this relation and it is, thus, not necessary to define it.

In this section, the schematic marks have been presented as a method to link data. These marks are defined in RDF and allow the identification of specific elements in XML schemas and in an OWL schema. Links have also been defined between schematic marks in order to automatically populate the ontology from XML documents. Figure 3 and 4 show two snapshots of the tool Protégé [PRO 08] which allows to edit an OWL ontology. It shows the different classes "Cell" and "Image" defined for the integration of data and their properties. These classes are directly used in our prototype to identify cells in video images (Figure 5).

Concerning the cell identification process, the mapping of XML schemas is manual, but the population is done automatic from the process. However, relationships that are not defined in RDF mapping rules cannot be used automatically for the population process. Consequently, some relationships between data have to be established by the patricians. For instance, the extraction of cell regions from an image cannot identify the kind of cells. In order to link the image regions to a certain kind of cell, this definition has to be defined manually.

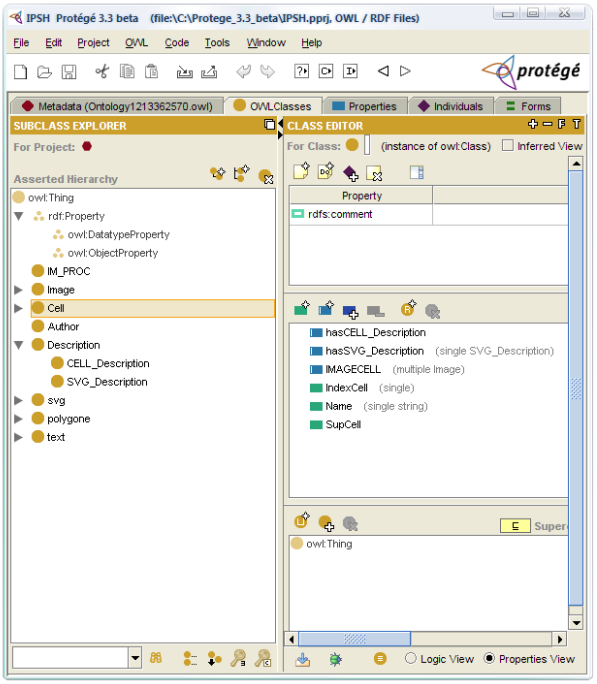

Figure 3. *This figure shows a snapshot of the tool Protégé that allows to edit an OWL ontology. In this snapshot, the properties of the concept "Cell" are shown.*

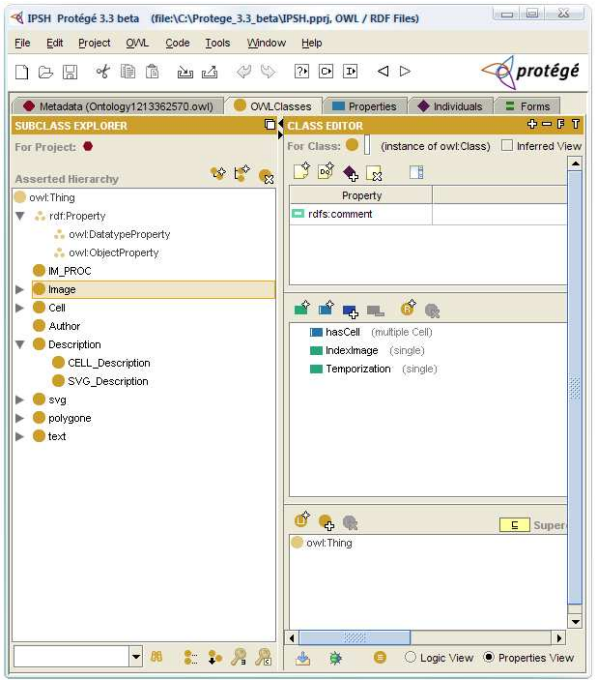

Figure 4. *The properties of concept are represented in green, such as "IndexImage" which is the identifiant of an image. The relation to an instance of another class which is a cell is represented in blue.*

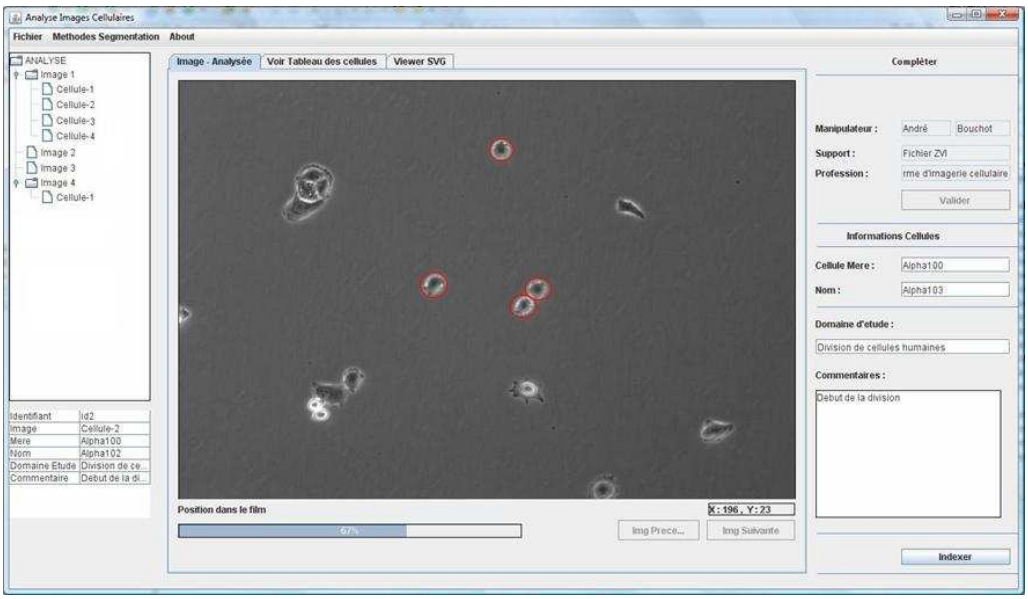

Figure 5. *This is a snapshot of our tool. The image in the center contains red circles that represent cells.*

5. Implementation

We have developed a tool (e.g. Figure 5) that allows to identify cells in video images. It allows also to index data with the help of an ontology. From this point, SVG data, cell data and video data are integrated in this platform. However, the method is flexible enough to add any other information that is relevant for the patricians. During the analysis process, all the detected cells found generate an individual of the corresponding concept Cell. The tree on the left of the prototype is generated from the ontology which contains all the data.

We are currently prototyping our tool using JENA. The tool runs on a server with Windows XP, the processor is a Pentium 3.2 GHz and 3.5 Go for the RAM. The process is carried out on a computer in the local network. To implement the framework we used the following tools: JENA (Semantic Web Framework for Java) [MCC 04] is used to build Semantic indexation in JAVA. In the development of processes, JENA helps us to handle RDF data. SPARQL [PRU 08] is a query language for RDF. SPARQL is used to manage RDF graph data like SQL request relational data.

6. Conclusion

In this paper, we have presented a data integration method based on the semantic marking of XML schemas and an OWL ontology. This method is applied for the combination of heterogeneous data coming from various services of a technical platform dedicated to cellular imagery processing. The data are at the same time SVG documents resulting from algorithms of image processing and specific business data describing work to be carried out or the results obtained. The use of an ontology makes it possible to associate a context of use with the resulting integrated schema. This context will be used for the development of a collaborative interface containing the view adapted to each user context.

ACKNOWLEDGMENT

The authors wish to thank C. Janny, A. Leclerc, P. Levitte and J. Tatibouet for their help in this project.

REFERENCES

[BAT 84 3] Carlo Batini, Maurizio Lenzerini: *A Methodology for Data Schema Integration in the Entity Relationship Model*. IEEE Trans. Software Eng. 10(6): 650-664 (1984)

[BER 00 4] Berstel, J., Boasson, L., 2000. *XML Grammars*, MFCS 2000: 182-191.

[BER 01 18] Berners-Lee, T., Hendler, J., Lassila, O.: *The Semantic Web*, Scientific American. (2001), 34-43.

[CAL 01 5] Cali, A., De Giacomo, G., Lenzerini, M., 2001, *Models for Information Integration: Turning Local-as-View into Global-as-View*, Proceedings of the International Workshop on Foundations of Models for Information Integration.

[CAS 07 30] Castano, S., Espinosa, S., Ferrara, A., Karkaletsis, V., Kaya, a., Melzer, S., Moller, R., Montanelli S., Petasis, G.: *Ontology Dynamics with Multimedia Information: The BOEMIE Evolution Methodology*. In Proc. of International Workshop on Ontology Dynamics (IWOD) ESWC 2007 Workshop - 7 June - Innsbruck, Austria, 2007.

[CRU 04a 33] Cruz, I. F., Xiao, H., Hsu, F., 2004.: *An Ontology-based Framework for Semantic Interoperability between XML Sources*, In Eighth International Database Engineering & Applications Symposium (IDEAS), 2004.

[CRU 04b 7] Cruz, I. F., Xiao, H., Hsu, F., 2004. *An Ontology-based Framework for Semantic Interoperability between XML Sources*, In Eighth International Database Engineering & Applications Symposium (IDEAS 2004).

[CRU 06 36] Cruz, C., Nicolle, C.: *Ontology-Based Integration of XML data*, Webist, Setubal, Portugal, pp. 30-37, 2006.

[FAA 04 32] Faatz, A., and Steinmetz, R.,: *Precision and recall for ontology enrichment*. In Proc. of ECAI-2004 Workshop on Ontology Learning and Population, Valencia, Spain, Aug. 2004.

[GAR 05 28] García, R., Celma, O.: *Semantic Integration and Retrieval of Multimedia Metadata*, *Proceedings of 4rd International Semantic Web Conference*, Galway, Ireland, 2005.

[GRU 08 27] Gruber, T.: *Ontology, to appear in the Encyclopedia of Database Systems*, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2008.

[HAN 02 25] Handschuh, S., Staab, S.: *Authoring and annotation of Web pages in CREAM*, Proceedings of the 11th international conference on World Wide Web, 2002

[KAH 01 24] Kahan, J., Koivunen, M.-J., Prud'Hommeaux, E., Swick, R.: *Annotea: an open RDF infrastructure for shared web annotations*, 10th International World Wide Web Conference (WWW 2001), Hong Kong.

[KLE 02 10] M. Klein, *Interpreting XML via an RDF schema*. In ECAI workshop on Semantic Authoring, Annotation & Knowledge Markup (SAAKM 2002), Lyon, France.

[LAK 03 11] Lakshmannan, L. V., Sadri, F., 2003. *Interoperability on XML Data*, In Proceeding of the 2nd International Semantic Web Conference (ICSW'03).

[MAR 07 31] Martin, D., Paolucci, M., Wagner, M.,: *Towards Semantic Annotations of Web Services: OWL-S from the SAWSDL Perspective*, In OWL-S Experiences and Future Developments Workshop at ESWC 2007, June, Innsbruck, Austria, 2007.

[MCC 04 15] Philip McCarthy (June 2004*), Introduction to Jena* [Online]Available:http://www-128.ibm.com/ developerworks/java/library/j-jena/

[OWL 08 19] Web Ontology Language (OWL), http://www.w3.org/2004/OWL/

[PRO 08 38] Protégé, *The Protégé Ontology Editor and Knowledge Acquisition System*. http://protege.stanford.edu/

[PRU 08 16] Eric Prud'hommeaux and Andy Seaborne, *SPARQL Query Language for RDF* [Online] Available: http://www.w3.org/TR/rdf-sparql-query/

[RAP 02 12] Pan, A., Raposo, J., Álvarez, M., Montoto, P., Orjales, V., Hidalgo, J., Ardao, L., Molano, A., Viña, Á., 2002, *The Denodo Data Integration Platform*, VLDB, Hong Kong, China.

[RDF 04 17] Resource Description Framework, *Model and*

Syntax Specification, W3C Recommendation (2004), available at http://www.w3.org/RDF/

[RUS 08 21] Russell B. C., Torralba, A., Murphy, K. P., Freeman, W. T.: LabelMe: *A Database and Web-Based Tool for Image Annotation*, International Journal of Computer Vision, Springer Netherlands, ISSN0920-5691 (Print) 1573-1405 (Online), Volume 77, Numbers 1-3 / mai 2008.

[SCH 03 23] Schroeter, R., Hunter, J., Kosovic, D.: *Vannotea: A Collaborative Video Indexing, Annotation and Discussion System For Broadband Networks*, Knowledge Markup and Semantic Annotation, Workshop, K-CAP 2003, Sanibel, Florida, October, 2003.

[SWR 04 29] SWRL: *A Semantic Web Rule Language Combining OWL and RuleML*, http://www.w3.org/Submission/ SWRL/, 2004.

[THA 08 37] Huynh Quyet Thang, Vo Sy Nam, *XML Schema Automatic Matching Solution*, International journal on Information Systems Science and Engineering, vo.l 4, number 1, 2008.

[THO 02 26] Thomas, C.: *Conceptual, Semantic, Syntactic, and Lexical Model*, 2002.

[URE 06 22] Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., Iravegna F.*: Semantic annotation for knowledge management: Requirements and a survey of the state of the art*, Journal of the Web Semantics: Sciences and Agents on the World Wide Web 4, Elsevier,14-18, 2006.

[WEL 99 20] Welty C., Ide N.: *Using the right tools: enhancing retrieval from marked-up documents*, J. Computers and the Humanities, 33(10)(1999) 59-84, 1999.

[TRI 08] TriG syntax, This document describes TriG, a syntax for serializing Named Graphs and RDF Datasets. http://www4.wiwiss.fu-berlin.de/bizer/TriG/