# Metamorphosis – A Topic Maps Based Environment to Handle Heterogeneous Information Resources

José Carlos Ramalho, Giovani Rubert Librelotto, and Pedro Rangel Henriques

Departamento de Informática – Universidade do Minho,
Campus de Gualtar 4710-057 – Braga – Portugal
{jcr, grl, prh}@di.uminho.pt

**Abstract.** Nowadays, data handled by an institution or company is spread out by more than one database and lots of documents of different types. To extract the information implicit in that data, it is necessary to pick parts from those various archives. To obtain a general overview, those information slices should be integrated. Different approaches can be followed to achieve that integration, ranging from the merge of resources till the fusion of the extracted parts. In this paper, we introduce Metamorphosis – a Topic Maps oriented environment that enables a conceptual navigation among heterogenous information systems – and we argue that Metamorphosis can be used to achieve, via Topic Maps, the referred semantic integration.

## 1 Introduction

Daily, a lot of data is produced by every institution or company. To satisfy the storage requirements, these organizations use most of the times relational databases, which are quite efficient to save and to manipulate structured data. Unstructured data (appearing inside documents) is stored in plain or annotated text files.

There is a problem when these organizations require an integrated view of their heterogeneous information systems. It is necessary to query/exploit every data source, but the access to each information system is different. In this situation, there is a need for an approach that extracts the information from those resources and fuses it. Usually this is achieved either by extracting data and loading it into a central repository that does the integration before analysis, or by merging the information extracted separately from each resource into a central knowledge base.

We use Topic to address the the problem of information integration mainly because it is the international industry standard – ISO/IEC 13250 – for semantic information integration and secondly because of its pure abstract nature (enabling the specification of every sort of ontologies). We are using successfully, for some years, this technology for classification and integration of documents in some use case scenarios [LRH03a, LRH04a].

L. Maicher and J. Park (Eds.): TMRA 2005, LNAI 3873, pp. 14–25, 2006.

In most semantic or knowledge based applications an ontology is composed of two parts. The classification structure or or semantic network and the catalog. The semantic network is composed of abstract concepts and their relations. The catalog is made of concrete information items. Throughout the paper when we refer to the term ontology we are considering the whole thing, the semantic network populated with catalog's information items.

However, the process of ontology creation is complex, time consuming, and it requires a lot of human and financial resources: it is necessary to specify the entire semantic network and all the information items that are going to populate it. In an Enterprise Information Integration scenario this can mean the manual extraction of many information items from several and different information sources.

To overcome this problem, we developed Metamorphosis. Metamorphosis is composed of several modules with different aims:

**Metamorphosis Repository (MMRep).** This is the central component and its purposes are the storage of Topic Maps (for the moment it imports and exports Topic Maps in XTM syntax). All the other components interact with MMRep (section 3 will detail this component).

**Topic Map Discovery (TMDiscovery).** TMDiscovery is a Topic Map driven browser and can be seen as a web interface to the MMRep (section 4).

**Topic Map Extractor (Oveia).** This component (still a prototype) automates the task of Topic Map harvesting; It enables the user to specify the extraction task and generates a Topic Map in XTM syntax that can be uploaded into MMRep (section 5).

**Topic Map Validator (XTChe).** XTChe is an implementation prototype of TMCL (*Topic Map Constraint Language*). It is not yet full integrated but in a near future TMDiscovery will have the power to change the Topic Map (insertion and deletion of topics), and then this module will ensure the preservation of the initial intended semantics (section 6).

In this paper we claim that with Metamorphosis the semantic integration of a set of heterogeneous information sources is possible to achieve. In order to achieve this we propose the following methodology:

1. Look at the information resources and decide how your conceptual view should look like;
2. Choose what information bits must be extracted in order to produce that conceptual view;
3. Specify the extraction task using Oveia;
4. Upload the generated Topic Map into MMRep;
5. Browse it with TMDiscovery and use this interface to access the information resources.

With this methodology the original information resources are kept unchanged and we can have as many different interfaces to access it as we want. We just have to create/generate/specify a Topic Map for each one.

In spite of its advantages MetaMorphosis should be used with some judgement. If your are dealing with frozen information sources, like historic databases, you will not have any problem. But if you are dealing with sources that are still changing you must be careful in defining the conceptual network, you should keep it above the level where the changes occur otherwise you will have to create a new Topic Map each time a change occurs.

The remainder of the paper is structured in the following sections: next section (sec.2) will introduce Metamorphosis, then a description of each module is presented with some detail (MMRep in sec.3, Oveia in sec.5, XTche in sec.6 and TMDiscovery in sec.4). Before the concluding remarks (sec.8) we present a real world case study to consolidate our proposal — *"Emigration Museum"* (sec.7).

## 2 Metamorphosis

The main idea behind Metamorphosis is close the gap between Topic Map technology and its users. Metamorphosis is being developed to become a Topic Map workbench easy to use and accessible to a common user (we are not there yet).

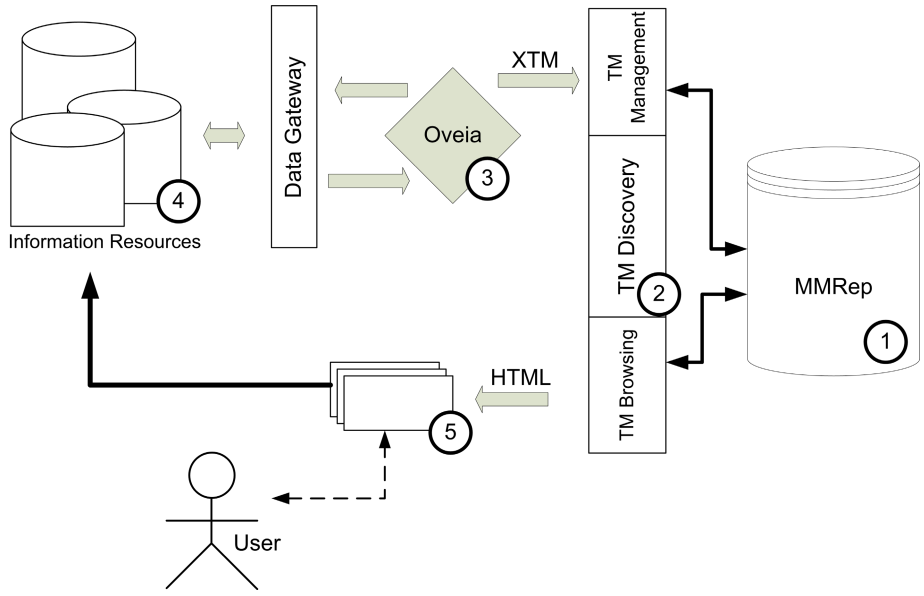


**Fig. 1.** Metamorphosis Functional Diagram

Figure 1 shows the usage scenario proposed in this paper. It illustrates some of the interaction between the system components, information resources and users.

1. MetaMorphosis Repository is the component that takes care of Topic Map storage and management.

2. TMDiscovery is the browser that allows users to navigate inside the Topic Maps stored in MMRep.
3. Oveia is a processor that eases the job of building topic maps. It implements some extraction mechanisms with which is possible to populate an ontology.
4. Information resources that we want to access.
5. Web interface driven by a topic map stored in MMRep that provides access to information resources.

Metamorphosis can be used to prototype web interfaces or to expose information systems on the web. To do this the user only needs to specify a topic map for each view he wants. Information integration is accomplished by concept integration in the topic map: to integrate two information systems we need to specify the two sets of concepts in the same topic map and specify the associations that will materialize that integration.

In the next sections we are going to discuss the main components of this workbench prototype: Metamorphosis Repository, Topic Map Discovery, Oveia and XTChe.

## 3 Metamorphosis Repository

Although XTM is a good format for interchange it is not so good for storage. When we refer to storage we are meaning the capability of storing a Topic Map and efficiently being able to query it. XTM is easy to process and for instance to translate it into another format. But querying XTM is complex. The Topic Map model is not hierarchical, every relation is materialized as a reference. Gathering all the information about a topic is very complex.

The obvious choice for storage is a database. For this case we had three options: an XML database [Bou05], an Object Oriented Database [Lea00] or a Relational Database. Since the Topic Map model does not match the XML model XML databases were discarded. Almost for the same reasons OO databases were also discarded. That left us with the relational model as the target for our storage solution.

The next step would be the specification of a Topic Map Relational Model. We have considered two approaches: look at the Topic Map Reference Model [Kip03, DN05] and derive the relational model from it or look at the XTM model and work from there. We decided to work over the XTM model and see if we could reach a model similar to the Topic Map Reference Model.

### 3.1 Data Model

First, we looked at the XTM model and raised the following subject list (and correspondent content model):

- $topicMap = (topic|association|mergeMap)*$
- $topic = (instanceOf|subjectIdentity|baseName|occurrence)*$
- $instanceOf = (topicRef|subjectIndicatorRef)$

- $subjectIdentity = resourceRef|(topicRef|subjectIndicatorRef)*$
- $baseName = (scope?|(topicRef|subjectIndicatorRef|resourceRef) + |baseNameString|variant*)$
- $scope = (topicRef|subjectIndicatorRef|resourceRef)+$
- $variant = (parameters, variantName?, variant*)$
- $parameters = (topicRef|subjectIndicatorRef)+$
- $variantName = (resourceRef|resourceData)$
- $occurrence = (instanceOf?, scope?, (resourceRef|resourceData))$
- $scope = (topicRef|subjectIndicatorRef|resourceRef)+$
- $association = (instanceOf?, scope?, member+)$
- $member = (roleSpec?, (topicRef|subjectIndicatorRef|resourceRef)*)$
- $mergeMap = (topicRef|subjectIndicatorRef|resourceRef)*$

After some exercise with the leaf nodes of this list we end with the following types that cover any element in a topic map:

| |
|---|
| $(topicRef\|subjectIndicatorRef\|resourceRef)$ |
| $(topicRef\|subjectIndicatorRef)$ |
| $(resourceRef\|resourceData)$ |
| $resourceRef$ |
| $baseNameString$ |

This result means that any Topic Map node can be represented with one of this five types. To store any of this five types we only need a triple: identifier, value and type. Consider the following example:

| Stored Values | | |
|---|---|---|
| **Id** | **Type** | **Value** |
| "TR982" | "topicRef" | "#University" |
| "SIR500" | "subjectIndicatorRef" | "http://www.uminho.pt" |
| "BNS32" | "baseNameString" | "U. Minho" |
| "RD444" | "resourceData" | "UM is ..." |
| "RR486" | "resourceRef" | "http://www.uminho.pt/students" |

This exercise enabled us to simplify the model and to reach the relational model showing in Fig.2.

With this specification we have implemented a Topic Map Repository that is the core component of Metamorphosis. In the following sections we will give some details about the integration of the other components with the repository.

## 4 Topic Map Discovery

Topic Map Discovery is an API that is being developed in order to work with the repository. For the moment it is composed of two parts: a topic map manager and a browser.

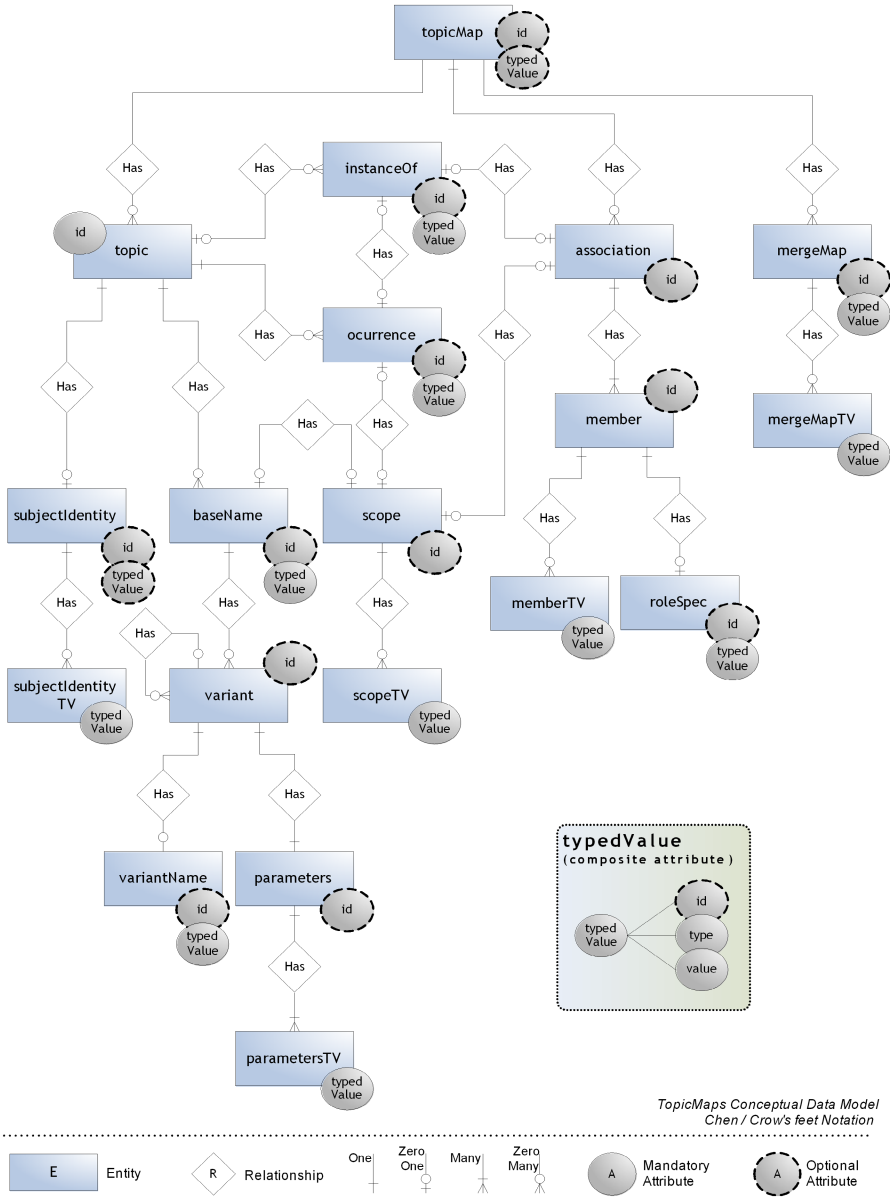


**Fig. 2.** Relational model schema

The topic map manager lets you upload and download topic maps in XTM syntax and delete a topic map from the repository (soon it will enable the user to edit stored topic maps).

The browser gives the user an interface to navigate inside any of the stored topic maps. So far we have developed the following interfaces:

**Topic Maps** - is the browser entry point and shows a list of all stored topic maps.

**Ontology Index** - gives you a structured view of a topic map showing the abstract concepts: topic types, association types, occurrence types and association role types.

**Individuals Index** - lists all non-type topics in alphabetical order.

**Full Index** - lists all named topics.

**Topic View** - lists a subset of the available information about a topic; for the moment: the basenames, its type, all the associations it participates in together with the other members and their roles, internal occurrences and external occurrences.

**Association View** - lists the names associated with the association and all its descendants.

We say that TMDiscovery is still a prototype because it has no control over users, there are no log or login components. However it will be freely available in a trial basis in the next days.

## 5 Oveia

The ontology extractor, Oveia (more details in [LSRH04]), is based on ISO/IEC 13250 Topic Maps [BBN99]. Oveia extracts information fragments from heterogeneous information systems according to an XSDS specification and builds the topic map according to an ontology specified in XS4TM language [LRH03b, LRH04b]. The Oveia architecture is shown in figure 3 and it is composed of five components. The dataset extractor receives an XSDS specification, providing metadata about the physical data sources that will be used to query each source in order to get the data needed for the ontology construction, and generates the intermediate representation (called datasets), containing the data (in a unified representation) extracted from resources. The XS4TM processor takes as input these datasets and an XS4TM specification generating a topic map, in XTM syntax.

### 5.1 XSDS — XML Specification for Data Sources

Oveia supports the concept of extraction drivers. A driver extracts data from a data source and stores it in an intermediate representation, called datasets. XSDS language defines the transformations and filters over the data sources. XSDS gives precise information about each data source that should be scanned to extract topics and associations.

An XSDS specification has two parts: *datasources* and *datasets*. The first one defines the path to the physical resources. Each resource is defined in a <*datasource*> element. This element has a set of attributes that indicates which extraction driver will be used and provides values for the corresponding parameters.

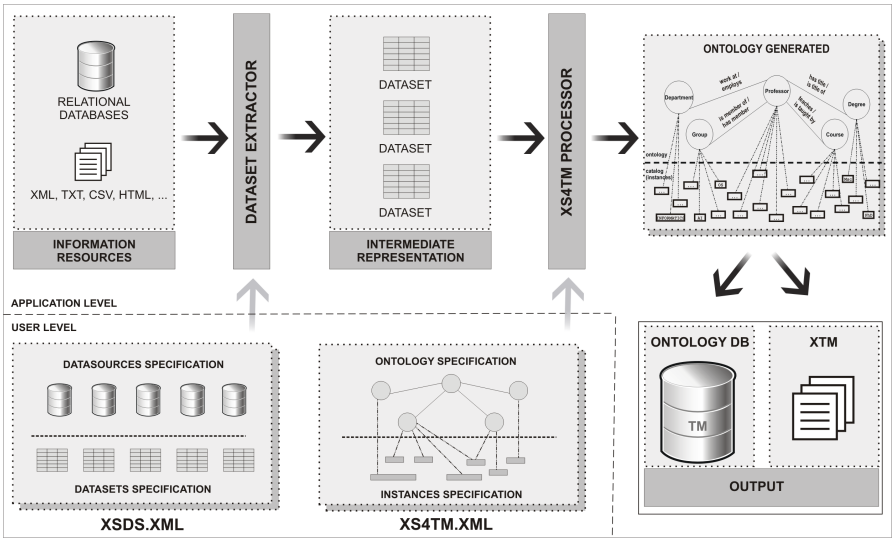


**Fig. 3.** Oveia Architecture

The second part of this specification is defined in a <*datasets*> element. It declares which data (record fields or DTD elements) must be extracted from each datasource. Each *datasource* can be used to specify the extraction of several *datasets*.

### 5.2 Datasets: Intermediate Representation

The *datasets* compose the intermediate representation that contains the extracted data from the resources. Each *dataset* has a relation to an entity in these resources and it is represented through a table, where each line is a record following the structure specified in XSDS. The *datasets* representation guarantees that Oveia sees an uniform data structure that represents all the participating resources.

The *dataset* declaration is composed of a query to extract the data from the resources. Each dataset has an unique identifier. This identifier will be used throughout the architecture to reference a particular *dataset*.

The fundamental idea is that all objects have labels that describe their meaning. For instance, the following object represents a member's category: $<1, PhD>$, where the string 1 is a identifier of this category, and $PhD$ is a human-readable label. The *datasets* are very simple, while providing the expressive power and flexibility needed for integrating information from disparate sources.

### 5.3 Dataset Extractor

The *Dataset Extractor* is a processor that scans the input data sources to get desired data into the *datasets*, in agreement with an XSDS specification.

The *Dataset Extractor* is composed of several extraction drivers (at moment, two), each one responsible for handling a specific type of source. The driver uses the appropriate technology to make the connection (e.g. JDBC, Java DataBase Connectivity, for databases, and an XML parser for annotated documents), and then the extraction of data is expressed in the query language adequate to the type of source in use: SQL will be used to extract information from a relational database while XPath will be used for the extraction in XML documents. Finally, the extracted data is stored in the *datasets*.

### 5.4 XS4TM — XML Specification for Topic Maps

XS4TM is a domain specific language conceived to specify the process of ontology extraction from information systems; in our case, from the dataset intermediate representation.

Looking at a topic map an ontology designer can think of it as having two distinct parts: an ontology and an object catalog (instances). The ontology is defined by topic types, association types, occurrence types, role types, etc. The catalog is composed by a set of pointers to information objects that are present in the resources and are linked to the ontology. So:

**Ontology.** The definition of the ontology requires in XS4TM the same effort as in XTM; it is necessary to specify every topic type, association type, occurrence type, ...;

**Instances.** The instances definition describes each topic and association that will be extracted from the intermediate representation; these will correspond to queries that will return lists of values; each value will turn in a topic.

The XS4TM Context Free Grammar is based in XTM 1.0 [PM01]. The *ontology* and *instances* elements have the same syntax as the *topicMap* element in XTM model.

The XS4TM language is intended to make the specification of Topic Maps extraction more flexible. However, the use of XS4TM is not much more difficult because this language is an extension of the XTM standard; it means the XS4TM DTD includes and augments the XTM DTD. In XS4TM, the ontology is specified like in XTM: with the same elements and attributes. So, if the designer knows XTM syntax, he does not need to learn another syntax to specify an ontology in XS4TM.

### 5.5 XS4TM Processor

This component uses the XS4TM specification and retrieves the information it needs to build the ontology from the *datasets*. It is an interpreter that takes advantage of the information organization in datasets (an internal universal representation for extracted data) and generates all the associations between the relevant topics according to XS4TM.

The XS4TM processor's behavior can be described in three steps: reads the the XS4TM specification and extracts from the datasets the topics and associations found; creates the topic map as an XTM file.

The status of this component is still prototype. The work that will integrate it with the repository is starting.

## 6 XTche – A Topic Map Semantics Validator

This component was already presented at Extreme Markup 2005 ([RLH05]). It will be integrated in Metamorphosis when editing capabilities become a reality. Until then it will remain a prototype as it is.

## 7 Emigration Museum: A Case Study

During the last centuries a huge number of Portuguese people (women and mainly men) left the country to go away to work abroad. Until the middle of twentieth century, the most important destination for emigrants was Brazil (an old Portuguese colony and a very large and rich country). Becoming rich, many of them came back and did notable things with real social impact; they constructed manor houses and palaces, schools, hospitals, churches, factories, and they developed the industry and commerce. As there are plenty of documents and evidences about those emigrants and the outcomes of there lives, a group of Historians in Fafe (a town in the North of Portugal) decided to create a virtual museum devoted to the Brazilian Emigration; after a first prototyped version(`www.museu-emigrantes.org`), we were involved in the conception of the information system.

The aim is to create a website that provides as many information as possible about each emigrant. The system should allow multiple navigation paths (offering various ways to handle the information) so that different views over the acquired knowledge are allowed. So, this museum on the Web should provide, not only data on individuals, but also knowledge about the social influence of their character and activities, in some geographical place at a certain date. To achieve that second, and main objective, it should be possible to cross data, exploiting the relations between the different information items (or units). Some interesting topics are: emigrant name; birth place and date; travel destination, departure and return dates, carrier; marital status; passport number; psychological profile; social or laboral event; industrial or commercial business; etc. Some important associations are: is; has; buy; creates; pays; offers; develops; etc.

However, as told above, the available resources, that should be exploited to extract the relevant data, are of many different kinds (official or technical records, literary documents, physical evidences, etc.), and are also available in different types of support: databases, annotated documents, and so on. For this case study we considered only three information sources: *travel diaries*, full of details written by the emigrant during the long (ship) trips; *biographical notes*, found in old almanacs, very rich in data concerning the character and social impact of the emigrant; *passport records*, obtained from the Portuguese foreign affairs bureau with factual data about travels. The first two are archived as XML documents (instances of two different document types), and the third one is a database.

In order to implement such an information system we could design a very large central repository, and impose that all the resources are consulted in order to extract the data to populate that huge database. Instead of that, we followed a completely different approach. We decided to use Metamorphosis to keep the data sources as they are and to generate a website where the visitor can start by accessing a topic and then navigate over the knowledge following the relations included in the underlying ontology.

Oveia was fed with: (a) the XSDS structural and physical description of the XML documents (*travel diaries*, and *biographical notes*), and the database (*passport records*) to be parsed to extract the relevant information bits to build topics; and (b) the XS4TM specification of the topic map to be built (notice that this TM corresponds to the ontology defined for the Emigration Museum). After some minutes, Oveia produced a 1,14MBytes (35588lines) XTM file containing a topic map with 1043 topics (instances of 25 topic types) and 1541 associations (instances of 32 association types). This topic map was then uploaded into MMRep and TMDiscovery allows users to browse the information accessing any item without needing to care about its origin. One of the pages, displaying the topic *emigrant*—that plays a role in 27 associations (of 12 different types)—is the most evident example of the knowledge integration achieved.

## 8 Conclusion

This paper describes the integration of heterogeneous information systems using the ontology paradigm, in order to generate an homogeneous view of these resources. The proposal is an environment, called Metamorphosis, for the automatic construction of Topic Maps with data extracted from the various data sources, and a semantic browser to navigate among the information resources.

Although developed for use in our main working area – XML documents processing applied to Public Archives and Virtual Museums – we are convinced that Metamorphosis can be applied with similar success in the general area of information system for data integration, analysis, and knowledge exploitation.

In the near future Metamorphosis will suffer several improvements. TMDiscovery will be able to edit topic maps. The inference engine behind the browser will be improved (for instance to give information about subtyping at any level). Oveia will be integrated in the management component of TMDiscovery. A friendly user-interface to write XS4TM and XSDS specifications is under development. Metamorphosis will be tested with new case studies, and we will conceive an easy and systematic way to verify the generated topic map against the actual sources and specifications. To assure the absolute correctness of this environment, each module should be formally validated.

As XTche specification language is based on XML Schema language, one of our next concerns is the implementation of the XTM-Skeleton-Extractor. The idea is to infer from the schema that specifies the constraints the basic specification of the Topic Map that we want to validate; this specification will be

the skeleton that the user can complete to obtain the XS4TM specification (the second Oveia's input).

## References


[BBN99] Michel Biezunsky, Martin Bryan, and Steve Newcomb. ISO/IEC 13250 - Topic Maps. ISO/IEC JTC 1/SC34, December 1999. http://www.y12.doe.gov/sgml/sc34/document/0129.pdf.

[Bou05] Ronald Bourret. Xml and databases. Website, Sptember 2005. http://www.rpbourret.com/xml/XMLAndDatabases.htm.

[DN05] Patrick Durusau and Steve Newcomb. Topic maps - reference model. ISO/IEC JTC1/SC34 Committee draft, February 2005. http://www.isotopicmaps.org/TMRM/TMRM-5.0/TMRM-5.0.html.

[Kip03] Neill A. Kipp. A mathematical formalism for the topic maps reference model. Draft paper submitted to ISO/IEC JTC1/SC34 Committee, October 2003. http://www.isotopicmaps.org/tmrm/0441.htm.

[Lea00] N. Leavitt. Whatever happened to object-oriented databases? *IEEE Computer*, August 2000.

[LRH03a] Giovani R. Librelotto, Jos C. Ramalho, and Pedro R. Henriques. ADRIAN – a platform for e-learning content production. In *Second International Conference on Multimedia and Information & Communication Technologies in Education*, 2003.

[LRH03b] Giovani R. Librelotto, Jos C. Ramalho, and Pedro R. Henriques. TM-Builder: Um Construtor de Ontologias baseado em Topic Maps. In *XXIX Conferencia Latinoamericana de Informtica*, La Paz, Bolvia, 2003.

[LRH04a] Giovani R. Librelotto, Jos C. Ramalho, and Pedro R. Henriques. ADRIAN E-Learning Content Production (creating online exams). In *VIII International Conference on Electronic Publishing*, Braslia, Brasil, 2004.

[LRH04b] Giovani Rubert Librelotto, Jos Carlos Ramalho, and Pedro Rangel Henriques. Extrao de Topic Maps no Oveia: Especificao e Processamento. In Mauricio Solar, David Fernndez-Baca, and Ernesto Cuadros-Vargas, editors, *30ma Conferencia Latinoamericana de Informtica (CLEI2004)*, pages 451–460. Sociedad Peruana de Computacin, September 2004. ISBN 9972-9876-2-0.

[LSRH04] Giovani Rubert Librelotto, Weber Souza, Jos Carlos Ramalho, and Pedro Rangel Henriques. Using the Ontology Paradigm to Integrate Information Systems. In *International Conference on Knowledge Engineering and Decision Support*, pages 497–504, Porto, Portugal, 2004.

[PM01] Steve Pepper and Graham Moore. XML Topic Maps (XTM) 1.0. TopicMaps.Org Specification, August 2001. http://www.topicmaps.org/xtm/1.0/

[RLH05] José Carlos Ramalho, Giovani Librelotto, and Pedro Rangel Henriques. Constraining topic maps: A tmcl declarative implementation. In *Extreme Markup Languages 2005*, Montral, Canada, August 2005.