



HAL
open science

Aid to the Semantic Maintenance of the Web Site

Michel Sala, Pierre Pompidor, Danièle Hérim

► **To cite this version:**

Michel Sala, Pierre Pompidor, Danièle Hérim. Aid to the Semantic Maintenance of the Web Site. ICWI 2003 - IADIS International Conference on WWW/Internet, Nov 2003, Algarve, Portugal. pp.369-377. lirmm-00269659

HAL Id: lirmm-00269659

<https://hal-lirmm.ccsd.cnrs.fr/lirmm-00269659v1>

Submitted on 21 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

AID TO THE SEMANTIC MAINTENANCE OF THE WEB SITE

Michel Sala
LIRMM
161, rue Ada
34392 Montpellier cedex 5

Pierre Pompidor
LIRMM
161, rue Ada
34392 Montpellier cedex 5

Daniele Herin
LIRMM
161, rue Ada
34392 Montpellier cedex 5

LASER
Richter – av mer BP 9606
34054 Montpellier cedex 1

ABSTRACT

The objective of our work is to provide some aid to the maintenance of a web site. The webmaster would like to get a semantic follow-up of the users' browsing, but he only has at disposal a set of statistical tools that indicate the frequency of one page visits but on no account by semantic aggregation of concepts. In this article, after having presented the methodology to modelise a web page while taking into account the semantic aspects, we will define the semantic analysis part of the web site and we will finish by a description of the global ontology in XML with instance storage. From local ontologies defined by users' categories, we are extracting semantic information. These results allow to do a revision of the site analysis or of the site design.

KEYWORDS

Web service modeling, ontology, semantic web, web, web service discovery, XML.

1. INTRODUCTION

In this paper, we describe an architecture that is going to allow to help a webmaster to analyze the uses of his web site we consider. The server of the web site has a set of statistical tools at the webmaster's disposal; which indicate the visit frequency of each page. Such an architecture will permit to modelise the site semantically and therefore will enable us to give semantic results according to the various browsing. During the exploitation phase, the server of the web site provides the webmaster with a set of statistical data such as the frequency of one page browsing. During this whole phase previously described, the semantic content of the web site is never used.

In our work, we take strong hypothesis:

- our study must be separated from the site development tools,
- we do not care the site's design but only the pages semantic contain,
- we do not modelise the information attributes in the pages, but only the global semantic,
- the studied site can be totally static.

Those different hypothesis neither allow us to use existing tools as ARENEUS [MER] or WebML [CER], more researchs based on dynamic sites [ATZ] [FER] [YAG].

Although at the time of the analysis the "customer" formalizes his need in informational and functional terms, he knows the semantics subjacent to every page of the site and even to every content of the page. By the word "semantics", we imply the meaning given to one page or a part of the page. We use the word "webmaster" to determine the analyst of the web site and the one who manages and maintains it.

Yet, the representation of the informational content presents many advantages like analysis a posteriori, of the browsing of the web sites' users. After having categorized the users, it is possible to know their semantic uses. While knowing the semantic uses and the content of the site, it is then possible to provide the webmaster with results of an abstraction level superior to those provided by the server's statistical tools.

In this paper, the application will be the software editor: PC SOFT [PCS]. One page of this web site contains the texts :

"We are software publishers : WEBDEV, AGL Internet, and WINDEV, AGL Windows, N°1 in France, and distributed in more than 50 countries." The semantics of this text is : Firm presentation

In our approach, all the results available to the webmaster are made by category of users. From the servers' logs, we define a set of categories of users obtained by selecting different users' similar browsing. For each category of users, we then know the pages browsed hence the semantics associated. In this article, we will not deal with the way the categorizations are calculated (works already exist in this domain such as : [BAU], [CHO]).

Example :

- a category of user is going to be interested only in the sub part « Product » of the site,
- another category is going to be interested in the sub part « Training offer » of the site.

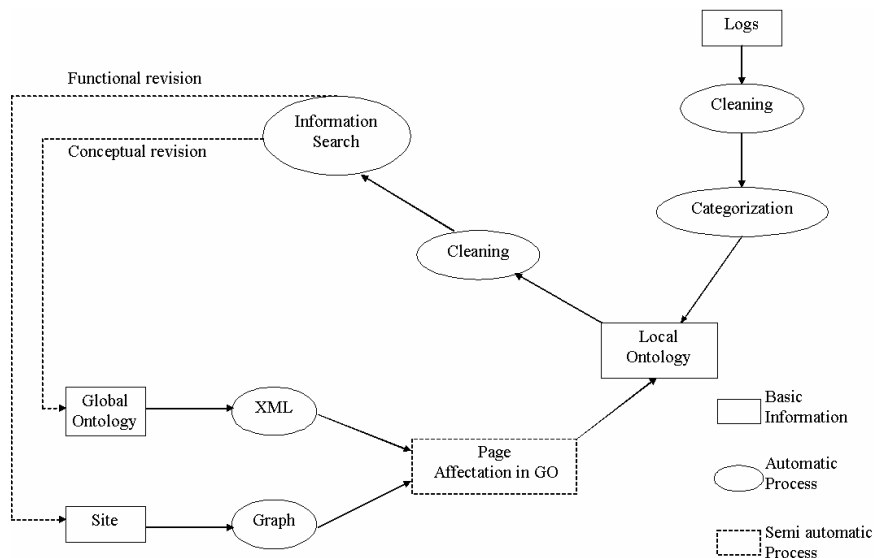


Figure 1. Global process

Figure 1 represents the global process used in the framework of our work. As prerequisites, we have a tour disposal the knowledge of the site as well as that of the semantic concepts used (called global ontology, described in paragraph 2.2). From the site initial conceptual analysis, a web page assignment is associated to the global ontology (described in paragraph 2.4). From the web server logs and users', we can define a local ontology associated to each category of users (defined by paragraph 3.2). According to every Local Ontology, our purpose is to be able to provide the webmaster with a series of semantic information, which will allow him to revise the analysis or the design of this site.

2. DESCRIPTION OF THE SITE INFORMATION

2.1 Analytical part

As a prerequi site, we take for granted that the initial stage of the web site analysis has been archived while using a traditional analysis method (Merise or UML [BOO]). In this case, we have a conceptual model of data (or class diagram) which represents the set of the concepts used in the web site with the inter-concept

relations (generalization, aggregation, association).

In the case where this analysis has not been achieved and/or to validate it, the following step describes a more thorough survey of the web site. This observation permit to analyze the informational content of the site as well as its structure and its dynamics. This survey permits to identify :

- the domains of information (manipulated concepts),
- the automatic generation of content via a database; we will then have to define the manipulated attributes of the database and their semantics,
- the semantics of the forms basing on the titles of the data capture areas, the unwinding lists...
- the hypertext links.

Example :

In figure 2, we describe the class diagram of the subpart Product of the PC Soft site.

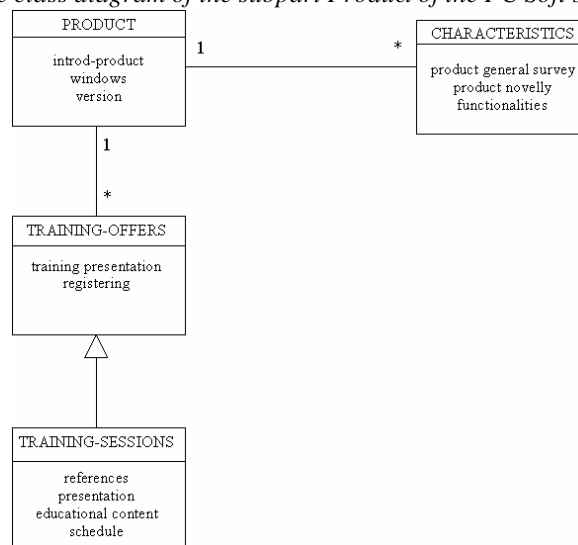


Figure 2. Extract of the class diagram of the PC Soft site around the PRODUCT class

We are going to use this data concept modelisation to extract the concepts manipulated in the site and assign each web site page to one or several concepts.

2.2 The notion of ontology

The scientific community has shown great interest in the concept of ontologies [GRU] to represent knowledge but also to make their sharing and their reusing easier. Among all the definitions given in literature, the most quoted one is from Gruber : "An ontology is an explicit specification of a conceptualization, i.e. a description of a part of the "world" in terms of concepts and relations between these concepts".

Our research works have led us to the development of ontologies in various domains :

- the Chimere project [SEG] using the ontologies as references for the questioning of several sources of heterogeneous data whose structure is not known a priori,
- in the domains of the didactics [HER]
- to help a scientific discovery by a researcher in experimental sciences [SAL].

Ontology plays a key role in the representation and the use of knowledge. It provides a coherent definition of the vocabulary used to represent knowledge, but it doesn't limit itself to a simple list of terms; it must also provide the semantic interpretation of these terms.

The elaboration of an ontology undergoes the following phases :

- - to delimit the domain of interest and the level of abstraction to describe it,
- - to define the vocabulary specific to the knowledge of the domain,
- - to modelise the knowledge in terms of taxonomy of concepts and individuals and in terms of the relations between them.

To define an ontology specific to the PC-Soft web site, we have following the different steps. In figure 3, we give a short extract of this ontology representing the concepts linked to a product.

Example : if we take the PC Soft site Product type, it can be described as follows: complexType Type_Products

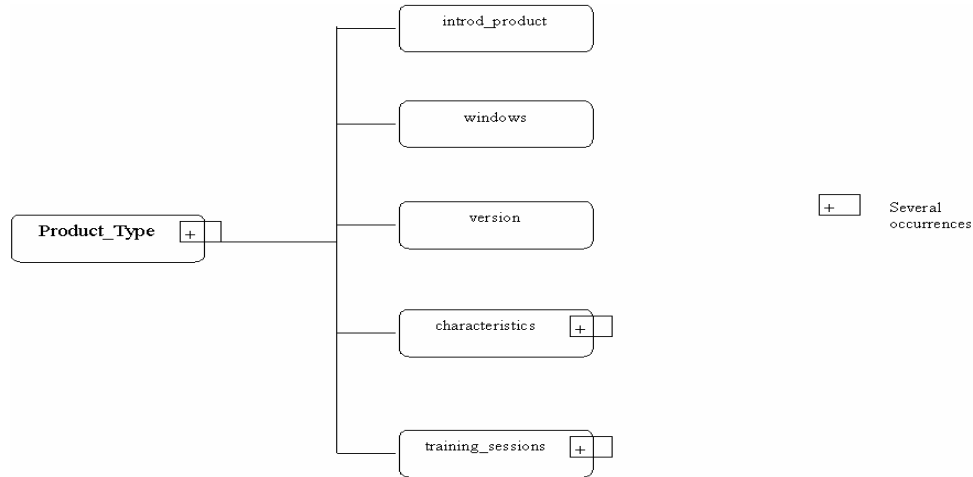


Figure 3. Extract of the ontology linked to the Product_Type

In our example, we can notice that node “Product_Type” can be split into son nodes “introduction_products”, “characteristics”, “windows”, job_training_offer” and “version”. The father of node “Product_Type” is the node “Product” that has itself father node “PC_Soft”.

We do not keep the initial modelisation (UML class diagram) although it is more thorough than the representation of the ontology chosen, since on account of a problem of transfer normalisation, we have chosen the XML format (defined in the following paragraph). The XML representation is tree-like (yet a node can be a sub-graph) and this formalism does not allow the representation of notions such as aggregation or heritage.

2.3 The representation of the ontology in XML

At the present time, many sites are still designed while using HTML. This "language" doesn't have any data type definition and doesn't permit to describe the web page from a semantic viewpoint. For this reason, we must modelise the information contained in the web site, then describe it, using an XML formalism [XML].

On the other hand, the XML language aims to provide a well-structured description of the present information on a site, contrary to HTML. The qualities of an XML document required are that it shouldn't include any syntactic mistake, and that if there exists a description of the document structure, it must be valid.

An ontology can easily be defined using the XML language, the normalization of it being achieved by the w3c. Different projects have been achieved like DAML (Darpa Agent Markup Language, www.DAML.org) and OIL (Ontology Inference Layer) [FEN]. Currently the OWL project (Ontology Web Language [OWL]) which is the fusion of the first two projects didn't lead to an official recommendation by the w3c.

If you go back to the example figure 3, its representation in XML can be written as follows :

children	<u>introduction_products</u> <u>characteristics</u> <u>windows</u> <u>job_training_offer</u> <u>version</u>
used by	element <u>Type_Pcsoft/products</u>

source	<pre> <xs:complexType name="Type_Products"> <xs:annotation> <xs:documentation> Product presentation section </xs:documentation> </xs:annotation> <xs:sequence> <xs:element name="introduction_products"/> <xs:element name="windows" type="Type_Windows" minOccurs="0"/> <xs:element name="version" type="Version_Type"/> <xs:element name="characteristics" type="Type_Characteristic" minOccurs="0"/> <xs:element name="job_training_offer" type="Training_Type"/> </xs:sequence> <xs:attribute name="name_type_product"/> </xs:complexType> </pre>
--------	---

2.4 Site page graph

The analytical part has enabled us to extract static knowledge linked to the web site; then we define a graph of dynamic link between the global ontology concepts and the web site pages. So, we analyze each page of the web site as well as the different elements in each (keys, pictures, scrolling menus, hypertext links...). Then, to each concept defined in the global ontology, we will associate the pages that concern them semantically. Thus, we must adopt a graph oriented type of data structure, having a root and no cycle. Each node represents a concept we associate the whole set of the pages (URL) referring to this node.

- However, in order to be exploited :
- the XMLSchema [SCH] description includes information that won't be exploited by the ontologies approach (in like multiplicities, definition type of the content...);
 - we still have to describe relations between the pages of the site and every element of the structure described.

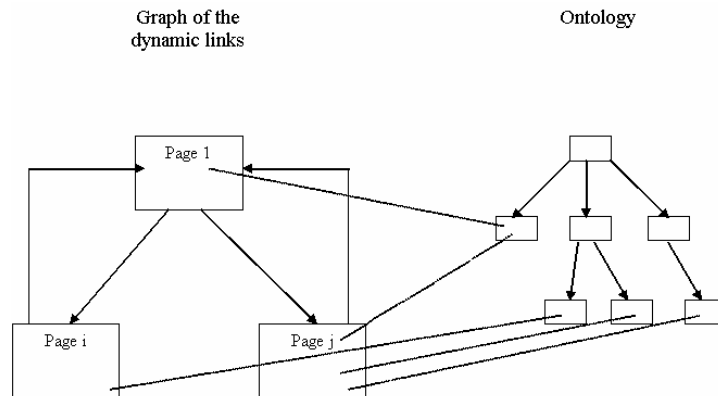


Figure 4. Graph of the dynamic links between the global ontology and the web site pages

We are going to use a representation that is limited to types of :

- <label> that contain the URL serving as a label, they have no attributes and contain no other tags,
- <node> that contains the description of an ontology concept. The node contain <label> tags (url associated to this concept) and can contain other tags <node> (concepts descending from this very concept). They can have a type attribute equal to the node reference : the node being normally described at the first occurrence of the reference, and described under reduced form (not closing tag) by its only reference to the other occurrences of it.

Using this formulation, we must however make sure that the node reference isn't found among the node's descendants, otherwise would generate a cycle.

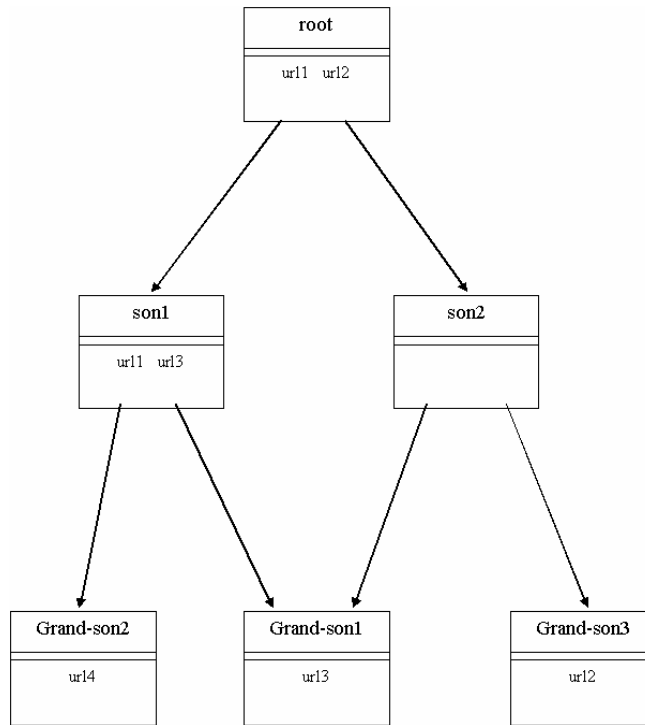


Figure 5. Example of nodes and links

In XML, we can represent the node son1 :

```

< node id="son1">
  < label >url1</ label >
  < label >url3</ label >
  < node id="grand-son1"
    < label >url3</ label >
  </ node >
  < node id="grand-son2"
    < label >url4</ label >
  </ node >
</ node >
    
```

Once the global ontology described and the links between the web pages and the ontology defined, we can work on the extraction of knowledge by categories of users which will allow to define the Local Ontologies and implement the ontology revision phase.

3. EXTRACTION OF KNOWLEDGE BY CATEGORY OF USERS

3.1 The notion of category of users

To get the categories of users, we start from the information contained in the web server transaction logs. They hold different information such as the date, hour, IP address, page browsed once these logs extracted from the server, we have a cleaning phase that enables to cancel the browsing made by mistake (for instance a useless search from a engine research [GOL]). We can define the user's oriented graph, but first for all, we will take into account only the pages whatever the order. All the same, we can determine how long the browsing lasted page by page (after cleaning all aberrations), which could eventually make the categorization sharper.

From different statistic tools which will not be detailed here, we get a set of categories of users that have more or less identical browsing of the site mentioned above. For a set category of users, we get a "local browsing" representing the usual browsing of the users of this category.

3.2 Extraction of the local ontology

A local ontology is a sub set of the reference ontology that we have called global ontology.

Example : in the PC Soft site, the set of pages devoted to training offers is a local ontology of the site global ontology.

So, for a category of users, we have described the set of the pages browsed. As the URL of the pages are stored in the nodes of the ontology concepts, we can find the nodes of the global ontology from these very pages. From the nodes selected, we extend this selection and search for the common ancestor to get a local ontology the algorithm of the local ontology extraction can be summed up as such :

- from the pages browsed by the set category of users, search for the nodes browsed,
- search for "linking nodes" to get an ancestor common to all the nodes previously selected.

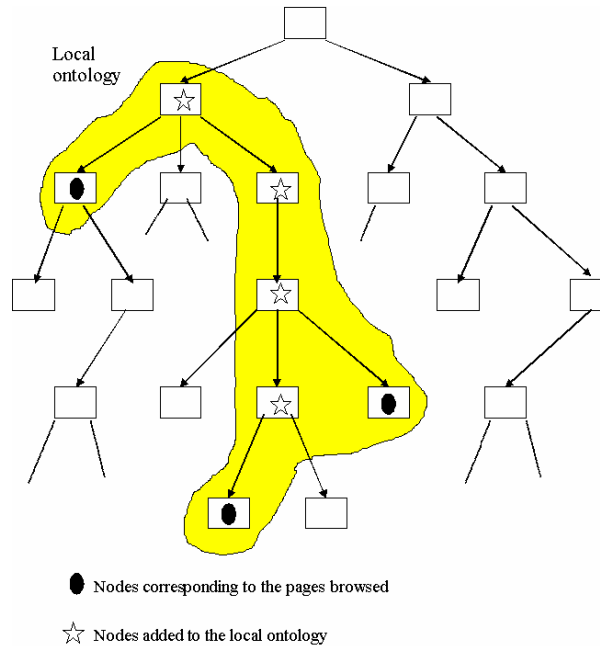


Figure 6. Extraction of a local ontology

4. GLOBAL ARCHITECTURE

The basic information of the web site falls into two types : the one linked to the site graph description, i.e. the pages used and the browsing graph as well as the site modelization in the form of a global ontology. Once these two analyses completed, we assign the referring web pages to each ontological node. Together with the site description, we categorize the different users from the actual browsing. For each category of users, with the help of the global ontology, we get a local ontology. Once we have collected all this information, we are now able to develop the mechanisms of information discovery and to propose to the webmaster to revise his site content. The global architecture and the process set up are represented in figure 6.

Once the local ontology defined for a category of users, we can clarify the part "discovery of the information" likely to help the webmaster maintain his web site. From the local ontology, we can obtain the set of the pages dealing with the concepts contained in this local ontology. Thus, we can analyse the pages to be browsed with the pages actually browsed. Once this set of : information obtained, we can move to the analysis phase to determine the reason for this difference.

To sharpen this analysis result, we have to develop a ponderation function which, from different information (such as two often a how long was a page browsed), will cancel the ontological nodes that are browsed nodes or wrong page browsing nodes (the user clicks on the link, gets to the new page, as is not interested in it, he goes back to the previous one).

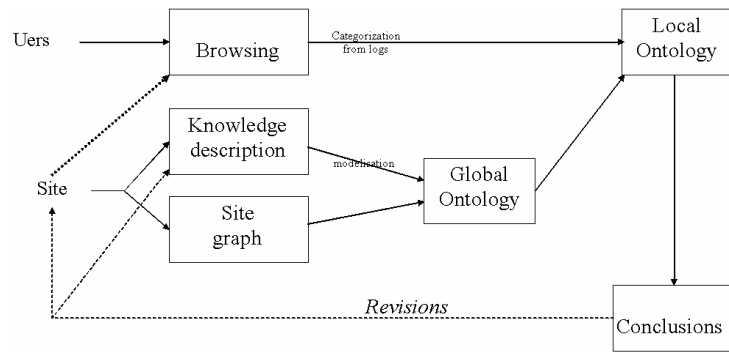


Figure 7. Global architecture and process set up

Once these calculations completed, we can detect the incoherencies between the pages actually browsed and those likely to be browsed, as they refer to the same ontological node (cf figure 8). Thus, for the pages really browsed, these can be two reasons:

- a wrong semantic analysis of the site pages and that has assigned some pages to one or several wrong ontological nodes,
- a bad ergonomic design that badly references or does not reference at all the link on a sub set of browsing (following repetitive maintenance additions for instance).

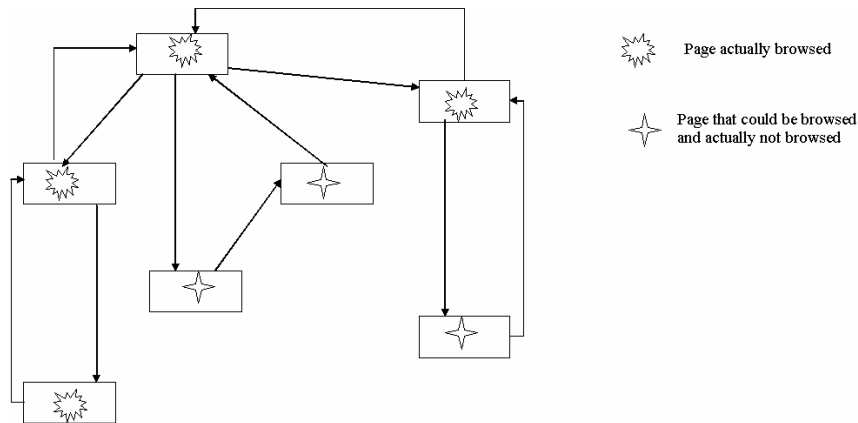


Figure 8. Detection of the pages that should have been browsed

5. CONCLUSIONS

At present, we have tested the modelling on two web sites, the former is the PC Soft publisher (presented in our example) which is a static site and of important browsing. The latter is the TIIM pole site (Technologies de l'Information, Informatique et Multimedia) [POL] which is managed in a dynamic way by using a database.

After the modeling phase (description of the global ontology, of the site graph, automation in XML, extraction of the local ontology, we aim at developing all the tools allowing to set up the phase of aid to revising the ontology. In that purpose, we will have to provide the web master with a series of explanations on the analysis completed. This explanation will include, for a given ontological node :

- a list of the pages actually browsed and, for each of them, the information analysed,
- a list of the pages that could be browsed, and, for each of them, the information analysed.

Example :

In the PC Soft site, the webmaster had designed an ontological node on employment and had assigned the pages devoted to the firm's job offers and demands. After analysis, it has been found that the category of users looking for a job never visited the pages containing the CVs of job-seekers.

Once the results analysed, the webmaster will be define the site global ontology by adding or putting

together one or several ontological nodes. Then he will dispatch the URL stored in the existing ontological node or nodes towards the new one or ones. This change will automatically generate the new XML representation of the global ontology as show in figure 9.

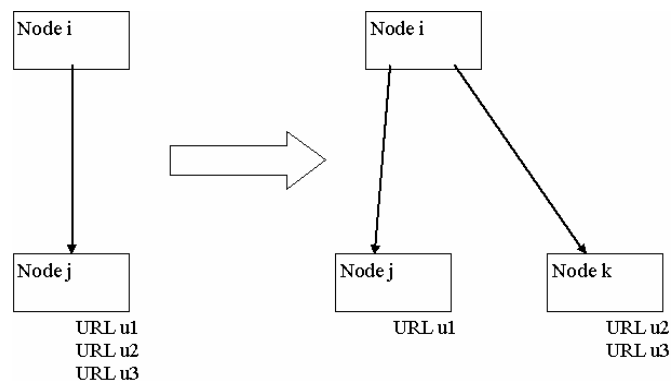


Figure 9. Revision of the global ontology

ACKNOWLEDGMENTS

We would like to thank P. Lauriac and F. Vabre for their involvement in this project.

REFERENCES

- [ATZ] Paolo Atzeni, Paolo Merialdo, Giansalvatore Mecca : Data-Intensive Web Sites: Design and Maintenance. World Wide Web 4(1-2):21-47, 2001
- [BAU] P. Baudracco, A-L. Beylot, C. Fleury, S. Monnier, M. Becker : Performance Measurement of the web server, Design of a first model. Research in official Statistics. Volume 1, number 1, 1998
- [BOO] G. Booch, J. Rumbaugh, I. Jacobson : The Unified Modeling Language User guide, Addison-Wesley 1998
- [CER] Stefano Ceri, Piero Fraternali, Aldo Bongio : Web Modeling Language (WebML): a modeling language for designing Web sites. WWW9 / Computer Networks 33(1-6): 137-157, 2000
- [CHO] H-K. Choi, O. Limb : A behavioural Model of Web Traffic. International Conference on network Protocoles, 1999.
- [FEN] D. Fensel, I. Horrocks, F. van Harmelen, D. McGuinness, P. F. Patel-Schneider : OIL An Ontology Infrastructure for Semantic Web In IEEE Intelligent Systems, Vol 16, N° 2, 2001
- [FER] Mary F. Fernandez, Daniela Florescu, Alon Y. Levy, Dan Suciu : Web-Site Management: The Strudel Approach. Data Engineering Bulletin 21(2): 14-20, 1998
- [GOL] J. Goldberg : www.goldmark.org/netrants/webstats
- [GRU] T. Gruber : A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5 : 19-220, 1993
- [HER] D. Hérin, M. Sala, P. Pompidor : Evaluating and revising browsing from web ressources educational, ITS 2002, Springer-Verlag LNCS, pp 208-218, juin 2002
- [MER] Paolo Merialdo, Paolo Atzeni, Giansalvatore Mecca : Design and development of data-intensive web sites: The araneus approach. TOIT 3(1): 49-92, 2003
- [OWL] www.w3.org/TR/2002/WD-owl-ref-20021112/
- [PCS] www.pcsoft.fr
- [POL] www.poletiiim.org
- [SAL] M. Sala, P. Pompidor, D. Hérin : A framework to review complex experimental knowledge, OOIS 2002, Springer-Verlag LNCS, pp 167-172, september 2002
- [SCH] www.w3.org/XML/Schema
- [SEG] M-S Segret, P. Pompidor, D. Herin, M. Sala : Use of ontologies to integrate some information semi-structured exits of pages web. INFORSID'2000, Lyon pp 37-55
- [XML] www.w3.org/XML
- [YAG] Khaled Yagoub, Daniela Florescu, Valérie Issarny, Patrick Valduriez : Building and Customizing Data-Intensive Web Sites Using Weave. VLDB 2000