2009

Web Observatories: Concepts, State Of The Art & Beyond

Nikos Nanas
, Laboratory for Information Systems and Services, Centre for Research and Technology - Thessaly, n.nanas@cereteth.gr

Manolis Vavalis
Laboratory for Information Systems and Services, Centre for Research and Technology - Thessaly, m.vavalis@cereteth.gr

John Koutsaftikis
Laboratory for Information Systems and Services, Centre for Research and Technology - Thessaly, j.koutsaftikis@cereteth.gr

John Koutsaftikis
,Laboratory for Information Systems and Services, Centre for Research and Technology - Thessaly, e.kelis@cereteth.gr

Elias Houstis
Laboratory for Information Systems and Services, Centre for Research and Technology - Thessaly, elias.oustis@cereteth.gr

WEB OBSERVATORIES: CONCEPTS, STATE OF THE ART & BEYOND

Nanas, Nikos, Laboratory for Information Systems and Services, Centre for Research and Technology - Thessaly, n.nanas@cereteth.gr
Vavalis, Manolis, Laboratory for Information Systems and Services, Centre for Research and Technology - Thessaly , m.vavalis@cereteth.gr
Koutsaftikis, John, Laboratory for Information Systems and Services, Centre for Research and Technology – Thessaly, j.koutsaftikis@cereteth.gr
Kelis, Elefterios, Laboratory for Information Systems and Services, Centre for Research and Technology – Thessaly, e.kelis@cereteth.gr
Houstis, Elias, Laboratory for Information Systems and Services, Centre for Research and Technology - Thessaly, elias.oustis@cereteth.gr

Abstract

The Web Observatories are becoming common Internet practice. They are web sites targeting a community of practitioners, scientists or generally individuals within the context of a focused organization. Their goal is to inform, educate, facilitate the interaction and boost the collaboration of community members. Various existing technologies can be deployed for this purpose. Still, their integration into a coherent informational and collaborative environment remains largely ad hoc. In this paper we attempt to elucidate the concept "web observatory" and identify its characteristics and practices.

Keywords: *World Wide Web; Information Filtering; Social Networking*

1 INTRODUCTION

Although the word ''observatory" is well understood it is difficult to formally define the term Web Observatory (WO) and maybe for some of us it is still unclear what a WO really is. Nevertheless, WOs have already proved themselves as rather powerful and effective instruments in several scientific, technological, educational and societal areas. They are surely emerging and we may claim that it is expected they will soon become part of our daily life.

Let us start with the following dictionary (http://www.m-w.com) definitions:

Observatory is

- a building or place given over to or equipped for observation of natural phenomena (as in astronomy and meteorology);
- an institution whose primary purpose is making such observations or
- a situation or structure commanding a wide view of its surroundings.

A meteorological observatory monitors meteorological conditions in a geographical area while an astronomical observatory surveys astronomical objects in the universe. Other types of observatories associated with the nature and various natural phenomena (e.g. ocean observatories, volcano observatories, etc.) also exist and in fact they are rather popular and extremely effective in achieving their goals and objectives. Likewise, health observatories (Hemmings & Wilkinson 2003) monitor and synthesise health related information over a region while language observatories (Mikami et al. 2005) survey language activities in the virtual universe over the World Wide Web.

The most common observatories are surely the ones concerning astronomy. They are wide-spread around the globe and range from tiny personal stations to huge international establishments. They all try to capture the weakest light from stars appearing far in the universe but they do not restrict themselves into a single objective and they usually have several goals. Besides satisfying pure scientific curiosity they also focus in introducing and popularizing the astronomy and space sciences, in providing testbeds for related scientific disciplines, in serving as data providers in associated repositories and databases, etc.

Likewise, a language observatory ''*tries to catch subtle messages of less spoken languages, as far as they appear in the virtual universe, and provides us such information like; How many languages are found in the virtual universe? What kind of character encoding schemes are employed to encode a given language?*"(Mikami et al. 2005). Similarly, a standards observatory (Anido et al. 2004) systematically and exhaustively follows the internet related standardisation process in a certain thematic area.

This natural context switch gave birth, more than 20 years ago, to a new type of observatories not directly related to natural phenomena. Since then, the number of such observatories kept increasing with several of them considered today very successful with significant overall impact (e.g. the Human Rights Watch). Right from their early days these observatories were organised around a web portal which initially served only as a dissemination media. As the time passed the role of this portal became more and more important for the observatory, sometimes to the extent that very few of their activities did not heavily involve the portal. This is essentially what we consider as Web Observatories and this is where our study focuses.

The challenge for these WOs is to line up their diverge ambitions with emerging trends in Information Technology, like the Web 2.0 and recent advances in information technology like Knowledge Management. Our paper consists an attempt to properly address this challenge. Its contribution is twofold; first to review and elucidate certain issues related to WOs and second to propose a framework and a prototype implementation that focuses on elevating the user's evolvement and interaction at a significantly higher level, by means of Web2.0 tools and techniques and recent advances in information filtering.

It is worth to mention the recently proposed concept of the *Virtual Astronomical Observatory* which can be viewed as a hybrid observatory that stands with one leg on the traditional notion of the observatories while it has its other leg on the WOs area we consider in this study. Based on this concept there exist many national and international large scale ongoing efforts like the European Virtual Observatory (Chen & Chen 2008) that collaborate through the *Virtual Observatory* (an international astronomical community-based initiative which operates under the auspices of the recently formed International Virtual Observatory Alliance). We can clearly see a shift in the paradigm where the astronomers now explore the digital Universe through practices, tools and instruments that are much closer to Information Technology than to Astronomy. This shift is by no means limited to Astronomy. It is already spread to similar thematic areas like Oceanography (e.g. World Ocean Observatory), where information technology tools coupled with traditional ones allow us to observe and analyse the ocean as an integrated, global, social system.

It is interesting to note here that a conventional observatory does not necessarily need to be neither big in size nor expensive to be build and maintained. This claim is well justified by the existence of a plethora of amateur astronomical observatories, small observatories that individual amateur astronomers have build since the 19th century for personal use and entertaining. Several of these observatories became very popular and are considered quite successful. We believe that neither WOs need not necessarily be expensive to be built and maintained, in particular in the Web2.0 era. We should finally mention that several WOs have been recently revised with Web2.0 in mind. The EU's Open Source Observatory is a typical example whose original WO was recently revised.

The rest of this paper is organised as follows. In the next section we present a short review of the characteristics and idiosyncrasies of a large class of selected WOs. Section 3 contains short descriptions of five recently developed WOs that share a common vision and identical objectives and which we have qualitatively evaluated. Next we focus on one of these that enjoys the desired characteristics of a next generation WO. Specifically in Section 4 we briefly describe the enabling Information Filtering core component of the system together with capability considerations and discussions on certain design and implementation issues associated with the next generation WOs and their methodological and technological frameworks.

2 BACKGROUND

We have identified a large collection of currently active WOs and collectively examined most of them. The detailed presentation of this examination is beyond the scope of this paper. Next, we briefly comment on certain of their characteristics. We should mention that the discussion that follows is associated with more than thirty WOs whose URLs together with short descriptions can be found at http://www.bibsonomy.org/search/web+observatories.

The general objective of these WOs commonly is to develop an apparatus for monitoring and evaluating the effectiveness of certain activities, the implications associated with various phenomena and trends, and the reactiveness of the individual persons or groups of individuals on such activities/phenomena. This apparatus will be based on both existing and emerging guides and toolboxes and will provide guidance, and practical examples, to all actors, who in various ways are affected by the processes of monitoring and evaluating the activities and/or phenomena under consideration. To develop such an apparatus one needs to address several issues and answer specific raised questions and challenges with the most important, to the targeted community, ones in our opinion being the following:

Selecting an observing design: How to construct useful observing logic and how to avoid typical modelling and implementation pitfalls? How to select the aspects of the issues that should be observed?

Collecting data: How to identify sources of data and how to plan for their gathering in the most efficient manner?

Analyzing data: Clarify how data are interpreted, how to avoid unneeded information and how to properly disseminate the most important information.

Each one of the WOs considered involves a set of main discrete tasks that include the following: collect information from other sources, measure indicators, publish newsletters and organize press releases, produce studies or surveys and prepare recommendations. Several are also the functions of a WO since it might operate as a learning network for members and participants; a single point of contact for external partners; an advocator for users of thematically or geographically restricted information; a coordinator of work across actors, etc.

An important issue is the way the data are collected and it turns out that it is a vital one. For the majority of the WOs considered in our study the data are hand picked. Several of them just integrate data obtained directly and automatically from particular databases which are remotely accessed and are distributed over the world. Few of them utilise agent technologies for web crawling and several produce additional data by processing existing ones through scientific workflows. It is quite apparent that unfortunately not many WOs effectively tried to involve users in the data collection process (Craven & Snaprud 2005) (beyond the level of poles and commenting). Several of these WOs that started with good sustainable prospects end up dead much earlier than anticipated (e.g. Observatory on the Information Society). There might be various and diverse reasons for that with user involvement in content creation being, to our experience, the most important.

Most of them are owned and operated by non-profit organizations (privately owned WOs do exist thought e.g. The Egnatia Motorway Observatory) with less than 5 employees and diverse geographical focus raging from regional/local (e.g. East of England Observatory, West Midlands Regional Observatory and Statewatch) national (e.g. Observatory for the Greek Information Society), continental (mostly European in fact e.g. European Internet Accessibility Observatory, European Observatory on Health Systems and Policies, European Quality Observatory, Internet Rights Observatory) and worldwide (e.g. Observatory on the Information Society,). The vast majority of the local observatories are bilingual (local official language and English) while only very few are language specific (e.g Observatoire des usages del'internet) and even fewer are in more than two languages. eGovernment, Information Society and the ICT sector are the main areas of interest while there also exist strong interest in eHealth, eBusiness, eLearning and education (Chabert et al. 2006), and eInclusion.

As already mentioned in the previous section, the virtual observatory (VO) concept is the astronomical community's response to recent IT advances and enabling technologies (Chen & Chen 2008). In particular, semantically enriched Web Services have been proposed in order to develop intelligent agents to represent: (1) users that submit requests (2) perform semantic matching in between users' requests and Web services registered within an agent platform, and (3) activate a serial of Web services.

Standarization activities in general naturally require a global view and an overall common perspective. Furthermore, it is really hard to get familiar with the learning technology standardization process and, even worse, to keep updated. A typical example is the Learning Technology Standards Observatory (Anido et al. 2004) produced by the European Committee for Standardization (CEN/ISSS WS-LT).

WOs are usually evolving. For example the International Observatory on End of Life Care whose objective is to provide information on palliative care services in Eastern Europe and Central Asia but aims in evolving it into a global information resource for the development of palliative care services.

Some additional notes about WOs characteristics can be summarized as follows:

They usually attract a significant amount of visitors while some are extremely popular (e.g. http://www.statewatch.org with 6 million hits per year). There is no clear dominant geographical focus, it is almost equally divided into regional, national European and international scopes. They come in various languages and they can be found in every continent (e.g. South African Cultural Observatory).

3 STUDY AND COMPARISON: REGIONAL INNOVATION POLES IN GREECE

Unlike Silicon Valley and other regions in the United States and parts of Europe that have evolved a working policy of public-private cooperation, Greece has been slow to adapt its legal and commercial framework to modern business and technology realities. In one of the recent efforts aiming to unlock the many technical achievements developed in Greek universities and research and development centers, Greece's General Secretariat for Research and Technology has established what it calls "regional innovation poles" in Central Macedonia (administered from the city of Thessaloniki), Western Greece (Patras), Western Macedonia (Kozani), Crete (Heraklion) and Thessaly (Volos) which let us associated to the acronyms CMACE, WGREE, WMACE, CRETE and THESS respectively.

Each one of these five Regional Innovation Poles has created a sectorial system of innovation, based on a small number of clusters or sectors (up to 3), and extended cooperation networks between the R&D labs, businesses, and technology intermediary organizations. They have also developed thematically and geographically focused web observatories that have the ambition to perpetually fertilize regional innovation in several thematic areas that include Energy (http://tw.innopolos-wm.eu) Quality and Business Excellence (http://www.quality-observatory.gr) Information Technology, Biotechnology, Medical Technology (http://www.i4crete.gr/), Textiles, Food and Biomedical Technology (http://www.rip-thessaly.gr).

	CMACE	WGREE	WMACE	CRETE	THESS
Technology Used	Joomla Javascript	Joomla v1.0 Javascript SMF Joomfish JEvents Calendar Freemeteo Plugin	ASPs Javascript	Drupal Javascript Localizer	Plone LinguaPlone GoogleMaps PloneGalleryView PressRoom PloneFormGen

Lay out	fixed-width 870 pixels 2 column	fixed-width 1045 pixels 3 column	fixed-width 1060 pixels 3 column	fixed-width 1000 pixels 2 column	Elastic-width 815 pixels 2 column

Table 38. The web technology used to develop and the layout characteristics of the Regional Innovation Poles considered.

We have evaluated the WOs associated with the five poles and our findings are summarized in Tables 1 and 2. Table 1 lists the main software tools and modules used for their development together with the most important layout characteristics. As it can be clearly seen, all are based on an open source Content Management System (CMS) coupled for some of them, with particular add-on products. It can also be seen that Javascript clearly plays a central role and there is enough wide diversity in the selection of the particular CMS. Some diversity can also be found in the layout characteristics. It should be pointed out here that none of the five truly follows the Web2.0 philosophy and practice. This is more apparent from Table 2 where we present the main findings of a qualitative comparison study. The effectiveness level for each one of the content and services items considered is given in terms of a scale from 1 (worst) to 5 (best).

		CMACE	WGREE	WMACE	CRETE	THESS
Content	Description	5	3	4	4	4
	Objectives	5	4	3	5	3
	Participants	5	4	3	5	4
	Actions	5	4	4	4	3
	News/Events	5	4	3	5	4
	Links	5	5	1	5	4
	Articles	5	4	3	4	2
	Multilingual	5	3	1	3	5
Services	Keyword Search	1	4	1	4	5
	Calendar	1	3	1	4	2
	User Login	5	4	3	5	5
	Forum	3	3	1	2	1
	Downloads	3	1	1	1	3
	Contact form	1	1	1	3	4
	RSS	1	3	1	1	5
	Newsletter	1	1	1	5	1
	Partner Search	1	4	1	2	2
Overall	Presentation	5	4	5	5	3
	Navigation	5	4	4	4	5

Table 39. Qualitative comparison of the Regional Innovation Poles considered in terms of content and services.

4 BEYOND THE STATE OF THE ART: COLLECTIVE FILTERING

In this section we will have a glance at a novice approach for developing next generation WOs which is part of an on going effort towards effective Web 2.0 personalization and adaptation.

4.1 Nootropia: the enabling Information Filtering (IF) core

Nootropia is a profiling model for adaptive content-based and/or collaborative IF. It is based on a weighted network (see figure 1), where, depending on the application, nodes can correspond to: features describing the content of information items, complete information items or users in a community. Links in the network represent the inter-relations between the node entities. A spreading activation process on the profile'�s network assesses how relevant an information item (or user) is to the profile. Inspired by the biological immune system, a self-organising process enables the profile to adapt to a variety of changes in a user's interests, or to changes in the information and social environment in general.

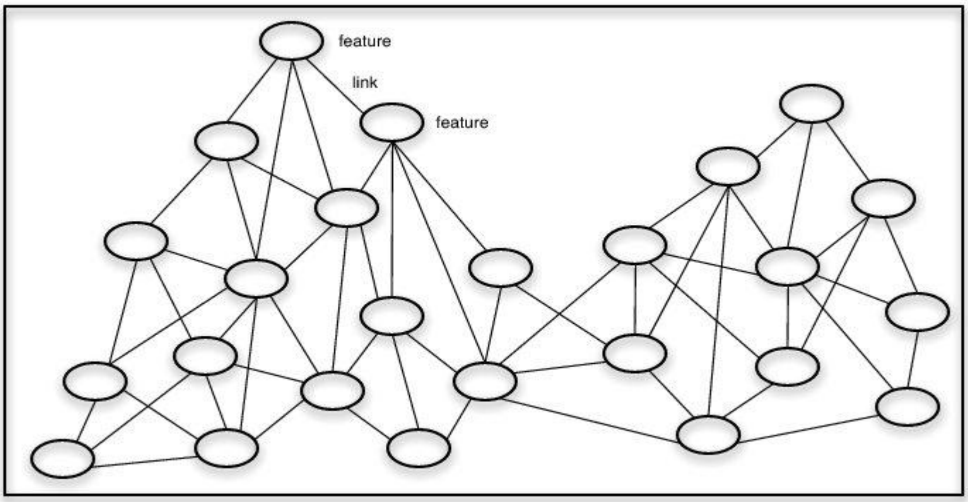

Figure 40. Nootropia: The basic network representation

More specifically, in the case of content-based filtering nodes in the network are features describing the content of information items. When dealing with textual information, these are terms extracted from documents and links in the network represent their statistical dependencies. Feature extraction from other media like image and audio is not as straightforward, but recent developments show that their content-based filtering will soon be feasible. Let us therefore concentrate on content-based filtering of documents, which will comprise most of the site's content. Nootropia will be used to built a user profile for each of the members in the Pole's community. It has the ability to represent a user's multiple interests and adapt to changes in them based on user feedback. Document evaluation assesses the distribution of relevance throughout a document's text and hence it can identify interesting documents and even pinpoint those portions of text that are more relevant.

4.2 System Architecture

The Innovation Observatory of Thessaly has two main modes of operation as these are architecturally depicted in figure 4. On the server side, the observatory's administrators can specify sources of information and in particular, RSS (Really Simple Syndication) feeds, which provide a daily influx of news articles related to the subject areas covered by the observatory (i.e. textiles, food/drinks and biofuel). The list of sources is easily altered and does not have to be predefined a priori. The news articles that the specified sources (RSS feeds) publish are indexed using the open source search engine Lucene. The indexing is done according to subject area, so search is possible within those RSS feeds that have been assigned to a specific subject area (e.g. textiles). The content of news articles is also parsed and analysed for extracting the most informative words, which are stored in a database.

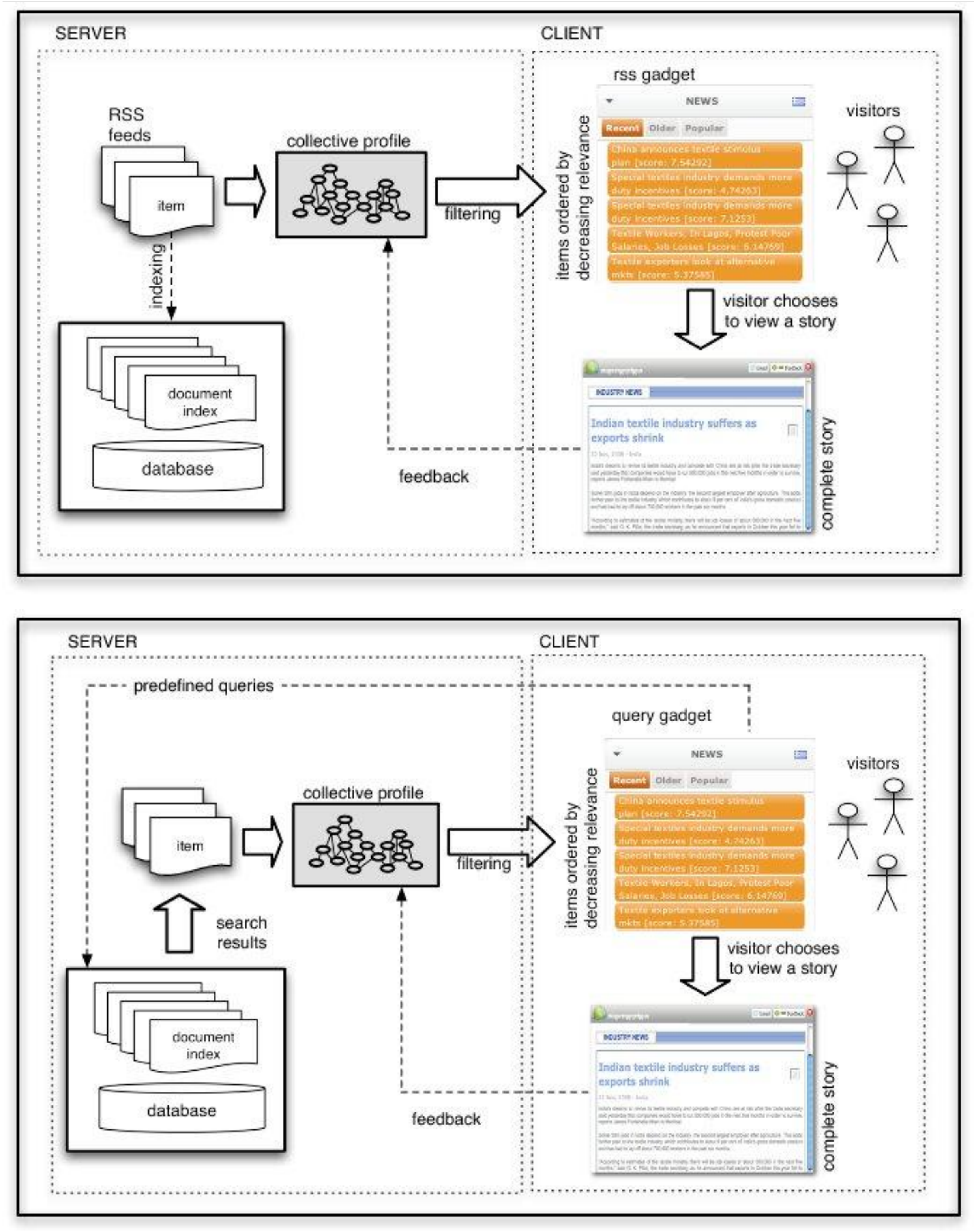

Figure 41. The architecture of the Innovation Observatory of Thessaly

On the client side, we used AJAX (Asynchronous Javascript and XML) to construct a responsive interface that is dynamically updated without the need to reload the web page. The observatory's web interface comprises two basic types of web widgets, or gadgets as these are shown in figure4. The first type of gadget is called ''RSS Gadget'' and it aggregates news stories from RSS feeds relating to a specific topic of interest (e.g. textile news, textile reports/statistics). The second type of gadget is called ''Query Gadget'' and it comprises a set of predefined queries, which are used for retrieving recently published, relevant news stories using the underlying search engine. For instance, one may use the queries ''cost'' and ''price'' to retrieve stories on the price of textile material from the corresponding RSS feeds. The observatory's administrators can easily modify the RSS feeds and the queries of RSS Gadgets and Query Gadgets respectively.

Back on the server side, each gadget corresponds to a collective profile based on Nootropia. The profile represents the collective interests of the observatory's visitors and accordingly evaluates the relevance of news. The result of this filtering process is a list of news stories ordered according to decreasing relevance. In the case of an RSS Gadget, the collective profile filters the news stories in the corresponding RSS feeds. Similarly, the collective profile of a Query Gadget filters the search results that the gadget's queries produce. A gadget is responsible for presenting the ranked list of news stories that its collective profile produces. Whenever a visitor of the observatory chooses to view one of the presented stories, the gadget's profile adjusts its structure according to the story's content. In this way the collective profile improves its representation of the visitor's interests and adapts to any changes in them over time. So even if the initial choice of RSS feeds, or of queries, does not accurately reflect the interests of the observatory's interests, the collective profiles ensure that what is finally presented to the observatory's visitors is adapted to their particular and changing interests.

4.3 Collective Filtering

The Innovation Observatory of Thessaly (publicly accessible at http://observatory.cereteth.gr) has been developed in the context of the Regional Innovation Pole of Thessaly (www.rip-thessaly.gr). Traditionally, the development of web observatories relies on the manual identification, categorisation and presentation of information relevant to a specific domain. However, this approach requires substantial human resources and expertise for keeping the web site's content up to date with developments in the domain of interest. In addition, even if such resources are available, there is no easy way to assure that the accumulated content satisfies the information needs of the observatory's visitors. To overcome these issues, we adopted a novel approach in developing the Innovation Observatory of Thessaly. An approach that relies on the dynamic enrichment, automatic categorisation and continuous adaptation of the observatory's content to the interests of it's visitors.

Figure 42. Innovation Observatory of Thessaly

The observatory's web site includes a series of web pages, but of interest to the current work are three of them, each one dedicated to one of the domains of interest (textiles, biofuel and food--drinks). Figure 5 depicts the web page that is dedicated to the domain of textiles. It comprises eleven widgets, each one dedicated to a specific topic category (Similar web pages were also built for the other two domains, biofuel and food-drinks). There are three types of widgets: static widgets, RSS widgets and query widgets. The static widgets are the simplest type and simply aggregate links that the observatory's administrator's have pre-defined. For instance, the ''COMPANIES'' widget contains a list of links to textile companies in the region of Thessaly. The second type of widget is an RSS feed aggregator. For example, the ''STATS'' widget aggregates information items from two RSS feeds dedicated to statistics and reports for the textiles industry. However, there are topics related to textiles, for which dedicated RSS feeds do not exist. For these topics, we used query widgets. We have identified a total of forty RSS feeds related to the textiles industry, including those which RSS widgets aggregate. All information items coming from these feeds are indexed using the Lucene search engine and become searchable for future reference. Query widgets use keywords to retrieve from the indexed feeds, information items relevant to a specific topic. For instance, the ''PRICES'' widget uses the keywords ''price'' and ''cost'' (and the corresponding Greek words) to retrieve recently published articles related to the cost of textile materials and products. This is of course just a rough method for classifying incoming information into predefined categories. However, each RSS widget and query widget is equipped with a separate profile, based on Nootropia. The widget's profile adapts the widget's contents to the interests of the observatory's visitors. In particular, whenever a visitor of the web page clicks on a title the corresponding article opens in a new browser tab and at the same time, it causes the profile's adaptation and the reordering of the widget's article list. Hence, the viewing patterns of the site's visitors provide implicit feedback that enables the profiles of the various widgets to track the changing community interests and to identify and present relevant information. So even if the initial classification of articles using keywords is not that accurate, the corresponding profile ensures that the most relevant items appear at the top of the widget's list. Of course, special care should be taken to avoid malicious behaviour, like for example a visitor that repeatedly clicks on a specific item's title. This can be avoided by only taking into account for the profile's adaptation a single click per title and per visitor IP address. Nevertheless, we do not expect such behaviour from the members of the focused community that the Innovation Observatory of Thessaly currently targets. Finally, we should add that the observatory comes with a simple interface for managing the links, RSS feeds and queries of the widget's.

5 SYNOPSIS

Web observatories are here to stay. Some of them may have been classified under different code names or keywords but they all share many characteristics in common and practices that we believe are worth to exploit and build upon. The component that forms the basis of what we call a WO is almost always a web portal, which allows for the inclusion of emerging Information Technology and Web2.0 techniques to enrich and enhance their functionality. We have collected a large set of WOs and briefly analyzed them. We also performed a comparative study of the five Regional Innovation Poles in Greece to outline the characteristics of their services, techniques and their presentation, forming the ground to propose a number of novel techniques. Information Filtering using the Nootropia model has been applied to the Observatory of Thessaly to provide personalised news services with the use of RSS feeds and a responsive and customizable innovative interface. Following current trends, Web2.0 techniques have been used to categorise news, events and search queries in RSS Gadgets which are dynamically updated and each adapted to the interests of the users that interact with them. This form of collective filtering keeps track of the demands of a community rather than a single user, with the items most useful to the sum being given priority and precedence over the ones that have received no implicit feedback. This process ensures that the most relevant items for a WO community are visible and that, should a switch in concept or a new interesting idea emerges, the WO will rapidly adapt to it.

The next generation WOs can go beyond being mere portals presenting information on a specific subject area. They can serve as online tools for Knowledge Management, enabling the creation, storage, dissemination and exchange of valuable information, relevant to the changing needs of a virtual community. They can get added value by tapping into the social network that emerges from the collaboration of individuals and into the rich content that their participation produces. Online trends like the Web 2.0 point away from the passive presentation of information selected and edited by some authoritative individual and towards the collective production and personalised delivery of information. Web sites become dynamic, continuously evolving and adapting to the community they serve. Adopting this new view towards WO poses new design requirements. The next generation WOs should adhere to the principles of both Knowledge Management systems and Web 2.0 sites.
We are currently developing a web application that will serve the information, communication and collaboration needs of an online community in general and a WO in particular. It will be based on a combination of private and collective profiles and provide a class of personalisation services that include: personalised search which evaluates the search results based on the user profile, recommendations based on the common interests of community members, expert finding that corresponds a user's request to the profiles of peers identifying those with desired expertise. We believe that the community of RIP Thessaly, or of other similar WOs provide an excellent opportunity to test this system in a real situation.

References

Hemmings, J. and J. Wilkinson (2003). What is a public health observatory? J Epidemiol Community Health, 57:324-326.

Mikami, Y. et al. (2005). The language observatory project (LOP). In Special interest tracks and posters Proceedings of the 14th international conference on World Wide Web, ACM Press, 10-14.
Anido, L., R. Rodriguez, C. Caeiro and S. Santos (2004). Observing standards for web-based learning from the web, Proceeding of the Computational Science and Its Applications, LNCS 3044:922-931.
Chen R.-S. and D.-K. Chen (2008). Apply ontology and agent technology to construct virtual observatory. Expert Systems with Applications, 34:2019-2028.
Craven J. and M. Snaprud (2005). Involving users in the development of a web accessibility tool. Ariadne Issue 44.
Chabert, G., J. C. Marty,B. Caron, T. Carron, L. Vignollet and C. Ferraris (2006). The Electronic Schoolbag, a CSCW workspace: presentation and evaluation. *AI Soc.* 20(3):403-419.