

Archiving the Greek Web

Charalampos Lampos¹, Magdalini Eirinaki¹, Darija Jevtuchova²,
Michalis Vazirgiannis¹

¹ Athens University of Economics and Business
Department of Informatics
Patision 76, Athens, 10434, GREECE
{lampos, eirinaki, mvazirg}@aueb.gr

² Vilnius University
Department of Mathematics and Informatics
Naugarduko str. 24, Vilnius, LITHOUANIA
darija@informaciniaprojektai.lt

Abstract. Web sites have become an increasingly important part of every country's information and cultural heritage. For this reason, Web archiving has become an issue for many national libraries. In this paper, we present a first attempt to archive the Greek Web. This project is divided in two parts; the first part concerns the collection of the majority of Greek Web pages. The second part focuses on the knowledge extraction from this archive, in order to classify it in semantically coherent clusters. Considerations concerning the criteria that should be set in order to characterize a Web page as "Greek" are discussed. A combination of IR and content mining techniques is applied in order to semantically characterize the collected content. We especially address the bilingualism issue arising because the content is written in both Greek and English. The collected Web pages are finally classified into meaningful clusters, facilitating the searching of the archive.

1. Introduction

The size and use of the World Wide Web has incremented exponentially during the past years. Web sites have become an increasingly important part of every country's information and cultural heritage. For this reason, Web archiving has become an issue for many national libraries. Such a Web archive provides a big amount of useful knowledge on many diverse areas. Due to the size of information which resides in Web sites and the dynamic nature of the most important of those (e.g. news portals etc), however, certain problems emerge concerning what should be included in a Web archive, in what format should it be stored and how often should this archiving happen. Many approaches exist, depending on the decisions made regarding each one of the aforementioned problems. In this paper we present a first attempt to archive the Greek Web. We address the bilingualism issue, since most of the Web sites include both Greek and (usually) English content. We apply Web mining methods to the archived content in an attempt to semantically characterize it. Using this semantic

annotation, a further categorization of the Greek Web content is performed using clustering algorithms.

The rest of the paper is organized as follows: in Section 2 we present some related national Web archiving projects. Sections 3 and 4 present in more detail our motivation, some preliminaries, as well as the methodology we followed in this project. In Section 5 we present some preliminary experiments, and we conclude in Section 6 with our plans for future work.

2. Related Work

The first efforts towards the creation of national Web archives began in 1996 by the Australian, Canadian and Swedish national libraries, as well as by the Internet Archive, a non-profit foundation. Since then, a big number of national libraries, as well as other institutions like universities have been involved in similar Web archiving projects. Moreover, the need for a common, nation-independent platform for supporting Web archiving and exchanging of information has led to the deployment of some European projects, such as NEDLIB [10]. An extensive index of most of those Web archiving initiatives may be found at [16].

There exist two main approaches for archiving the Web. The first one is to select manually a number of sites (usually a few hundred) and choose a frequency of archiving them. The Australian [4] and the Canadian [14] national libraries have followed this approach. The second approach is based on automatic selection by using Web crawlers. This was the approach followed by the Swedish national library [2], the Internet Archive [13], and lately from the French National Library [1].

The use of crawlers enables the archiving of a much wider range of Web sites. However, as this is an automatic process, decisions concerning the selection of Web pages to be archived, the frequency of the updates and the additional information that should be stored should be made. Issues such as the deep or invisible Web still remain open to discussion.

3. The Greek Web archive project

Influenced by the aforementioned approaches, we decided to create an archive for the Greek Web. This work is divided in two parts; the first part concerns the creation of an archive containing as many Web pages as possible. The second part focuses on the knowledge extraction from this collection of Web sites.

An estimation concerning the size of the Greek Web can be performed if we take into account the fact that there exist more than 60.000 Web sites registered in the .gr top-level domain. What should be characterized as “Greek” Web, however, is not solid since, for example, many Greek Web sites are hosted under the .net, .com, or .org top-level domains. Our effort focuses on archiving not only .gr Web sites, but all the Web sites that are Greek-oriented instead.

Since the amount of data collected is very big, it is necessary to somehow categorize it. For this purpose, we apply a combination of IR and data mining

techniques in order to characterize the Web content. This semantic characterization is then used in order to group the Web sites in thematic clusters. These clusters can subsequently be used to accelerate the search in the Web archive and enable the keyword-based search without human intervention. Our focus on this paper is mostly on the IR methods employed for indexing and further processing the archive, whereas updating/versioning issues remain open for further work.

3.1 The Greek perimeter

On the Web, it is difficult to define the exact limits of the Greek Web. The criteria that should be used in order to characterize a Web site as “Greek” are not solid, and most of the solutions seem ambiguous. We intend to archive any Greek-oriented Web page and not only the ones that are written in Greek or belong to the .gr domain. Therefore, the main criteria we use are:

1. *The domain name*: Most of the Greek Web sites belong to the .gr top-level domain. However, this is not inclusive, since many Web sites have .com, .net, or .org suffixes, depending on the kind of organization or company they refer to.
2. *The Greek language*: The use of the Greek language is a fundamental criterion for assuming that a Web page should be characterized as “Greek”. However, many Web sites are written in both Greek and English languages, and some of them are totally written in English language. Consider for example the Web pages of a University that provide information to the incoming Erasmus (exchange) students. These Web pages do not include content written Greek.
3. *“Greek” content*: A Web site may not have the .gr suffix in its name, or may not be written in Greek. However, it may contain information closely related to Greece. For example, there exist many Web sites written from Greeks that live abroad and aim at creating Greek communities or for “advertising” Greek culture and Greece in general abroad. Such sites can be tracked by examining their content (i.e. whether they include any of the words “Hellas”, “Greece”, “Greek”, etc.)

4. Methodology

The system that performs the archiving consists of three main components: the *Web Crawler*, the *Content Manager* and the *Clustering Module*. The Web crawler searches the Web using the aforementioned criteria in order to gather as many “Greek” Web pages as possible. The collected URIs are stored in a database along with the date and time the crawling was performed, to enable updating of the archive in the future. Some additional information such as the Web pages that point to, or are pointed by the URI can also be included for future use. Subsequently, a combination of keyword extraction methods is applied to the content of each Web page in the archive. Since the content to be processed is written in both Greek and English, an additional translation step is needed before concluding on the most representative keywords for each Web page. These keywords are also stored in the archive. Finally, we use this knowledge, to categorize the Greek Web into semantically coherent clusters. The

architecture of the Greek Web archiving system is depicted in Figure 1. The modules of this system are described in more detail in the following sections.

4.1 The Web crawler

Taking the aforementioned criteria into consideration, we proceeded in creating the archive of the Greek Web. The domain name (.gr) is the most important and reliable criterion to begin with. Moreover, the *charset* parameter of the HTML code of every Web page is examined. This parameter reveals the language in which the content of the page is written.

At this point we should stress that in this project our main interest was to gather as many Web pages as possible, regardless of their content. Any “Greek” page is of interest and should be included in the archive.

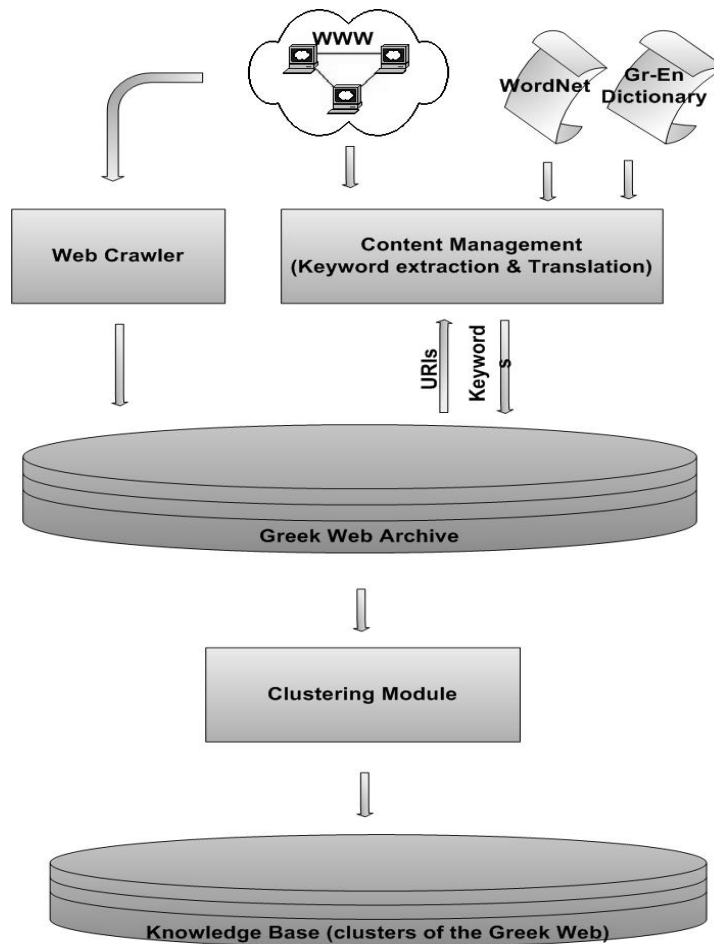


Fig. 1. Greek Web archiving system architecture

The crawler starts the searching from one or more Greek Web “hubs”. Big Greek portals which contain links to many other resources in addition to news, guides etc can serve as such hubs. The crawler follows all the paths that refer to Greek Web pages, taking into consideration the criteria described before. The search is exhaustive, following all the links from the subsequent pages, until a Web page with non-Greek characteristics is reached. Additionally, other sites, such as the Hellenic Resources Network Web site (<http://www.hri.org>), may be used. This site “provides Greek public policy institutions with improved access to the distinguished community of Greek academics, researchers and professionals living abroad, as well as the ability to communicate its positions and initiatives to a growing and dynamic part of the Greek Diaspora [...] enhances communication and cooperation between Greeks around the world by facilitating a more effective use of the Internet”. Therefore, it is a valuable “hub” to many other internationally located Greek Web sites.

At the end of the crawling process all the retrieved Web pages are stored in a database, along with additional information such as date/time, outgoing links of the page, links pointing to the page, etc. We chose to use a database instead of a file system since our initial objective was to gather the Web pages and further process them using clustering techniques. The architecture of the system, however, enables the storage of the retrieved Web pages in their original form as well.

4.2 Keyword Extraction

In order to be able to group the Greek Web into semantically coherent clusters, it is necessary to somehow characterize the content of the Web pages. The keyword assignment may be done either manually, or automatically. A manual assignment will result in accurate results, however the amount of data to be processed is huge, and therefore such approach is inefficient.

In our approach, we automatically extract keywords for characterizing each Web page. Most of the methods proposed for characterizing documents using keywords perform text mining in the document itself, following standard IR techniques. However, this approach proves insufficient when the document is a Web page, since it relies solely on the information included in the document itself, without taking into consideration the connectivity features of the Web [3, 6]. Moreover, it is difficult to extract keywords from Web documents that contain images, scripts, etc. Therefore, in many cases, the links that point to a document are used in order to extract information about the document itself. For this purpose, the text around the link to a page (called “anchor-window” [5]) is taken into account. This approach overcomes significantly the problems of the text-mining approach, since it takes into consideration the description others give to a Web page. A related effort capitalizing on link semantics appears in [7, 11].

In our project, we used a combination of the aforementioned techniques, along with a method based on the page links introduced in [7]. This latter method is based on the assumption that most of the Web pages include links that point to other pages containing topics that are of importance in the page’s context. Consequently, the keywords that characterize a Web page p are extracted using:

1. raw term frequency of p
2. raw term frequency of a selected fraction of the most important Web pages that point to p (inlinks)
3. raw term frequency of the Web pages pointed to by p (outlinks)

Hence, for each URI included in the archive, the Web page is visited, the textual content is isolated from structure data and after removing all the non-significant words (document indexing) using an appropriate stop-words list, the most frequent keywords are selected. A similar method is used for all the Web pages that are connected through a link with the Web page under examination. As far as the third method is concerned, we used Google's backwards link service¹ to select the first 20 pages pointing to the Web page. After experimentation, the anchor-window was set to 100 bytes.

4.3 Translation

The keywords that are extracted by applying the methods described above to English content need no further processing. But, what happens in the case of the Greek Web archive? It is evident that the archive contains pages written in Greek, in English, or both. Therefore, by using the keyword extraction methods without previously processing the keywords extracted may lead to wrong results, since, for example, a word that may appear in medium frequency, but both in Greek and English, will have a much higher frequency if the two frequencies are summed.

This is the reason why prior to selecting the most frequent keywords, all the Greek words are translated to English. However, this process is quite complicated. All words in the Greek language (nouns, verbs, adverbs) can be inflected. Hence, a first step includes the stemming and subsequent transformation to the nominative of each Greek word. For this purpose, we used the inflection rules of Triantafillidis Grammar [17].

After transforming all the Greek words to the nominative, they should be translated to English. However, this process is not straightforward since each Greek word has many English synonyms. Nevertheless, which one should be selected as the most appropriate translation heavily depends on the context of the Web page's content. A simplified approach would be to keep all possible translations, or a number of them. This would result in a big amount of keywords and would lead to inaccurate results. Another approach would be to keep the "first" translation that is given by the dictionary. However, since the translation is closely connected to the content of the Web page, the "first" translation is not always the best. For example the words "plan", "schedule" and "program" are some of the translations of the same Greek word ("πρόγραμμα"), however in the Informatics context, the word "program" is the one that should be selected.

In our project, we used a more complicated, yet reliable method. In order to decide which synonym is the most precise in the Web page's content, we take into account the sense of the rest keywords extracted for this page. Assuming that the set of keywords will be descriptive of the Web page's content, by comparing their

¹ http://www.google.com/advanced_search

semantics, we result to the best set of synonyms. The translation method is depicted in Figure 2. The procedure input is the set of English and Greek keywords ($En(D)$ and $Gr(D)$ respectively) of each document D . The output is a set of English keywords K that “best” characterize the web page. Let $En(g) = \{\text{english translations of } g, g \in Gr(D)\}$ and $Sn(g) = \{\text{Wordnet senses of keywords in } En(g)\}$. For every translated word’s sense (as defined by Wordnet [18]), the algorithm computes the sum of the maximum similarity between this sense and the senses of the remaining keywords (let $WPsim$ denote the Wu&Palmer [19] distance between two senses). Finally, it selects the English translation that has the sense with maximum score.

```

Procedure translateW(Gr,En)
1.  $K \leftarrow \emptyset$  ;
2. for all  $g \in Gr(D)$  do
3.   for all  $s \in Sn(g)$  do
4.      $score[s] = 0$ ;
5.     for all  $w \in En(D) \cup Gr(D) - \{g\}$  do
6.        $sim = \max(WPsim(s, Sn(w)))$ ;
7.        $score[s] += sim$ ;
8.     done
9.   done
10.  $s_{max} = s'$ ; ( $score[s'] = \max(score[s]), s \in Sn(g)$ )
11.  $K \leftarrow e, e \in En(g), e$  contains  $s_{max}$ ;
12. done

```

Fig. 2. The translation procedure

Empirical results have shown that this algorithm works well in practice and assigns the most relevant synonym to each Greek word (some examples are included in the Appendix). This module is part of the SEWeP system [8,15].

4.4 Clustering

At the end of the keyword extraction process, each Web page included in the archive is characterized by a set of English keywords that are descriptive of its content. This knowledge is further used to classify this content into semantically coherent clusters. This clustering process enables the automatic “segmentation” of the Greek Web. This segmentation augments and accelerates the searching of the archive. Instead of giving the exact URI name, the user can either perform keyword-based search, or cluster (thematic category)-based search. A further step would include the creation of a user interface to the archive that enables the user to search only the Web pages that fall under a specific category, and not all the contents of the archive.

In order to cluster the archive’s contents, the user may select to use K-means or DBSCAN [9] algorithms. Prior to applying one of these algorithms to the data, a preprocessing phase is needed in order to compute the TF_IDF space and calculate the coordinates of each document, which can further be used by the K-means and DBSCAN algorithms. The Clustering Module generates a label for each created cluster, taking the cluster centroid into account. The system also integrates a Cluster

Validation sub-module, in order to evaluate the quality of the created clusters. Since no a-priori structure of our data was known, we used relative cluster validation criteria [12], including the Dunn index, modified Hubert Statistics and Davies-Bouldin index for this purpose.

In the Web Archiving context, the input data size is very big. Therefore, an approach based on clustering the entire data set is inappropriate due to memory restrictions. To address this problem, the most prominent solution seems to be sampling the dataset, clustering the sample and classifying the remaining documents to the created clusters.

5. Experimental Evaluation

Using the system described above, we created a Web archive containing up to 300.000 distinct URIs belonging to “Greek” Web sites. The crawling was performed in September 2003 and lasted (along with the keyword extraction and translation) 2 ½ days. We used an initial seed of 3 Greek portals (*in.gr*, *pathfinder.gr*, *flash.gr*). The criteria used for characterizing a Web page as “Greek” were, as described before, the .gr top-level domain name and the Greek content. Since versioning is a fundamental feature of a Web archive, we also kept date/time details. The crawler used was a modified version of the crawler used in the SEWeP [15] and THESUS [11] systems. Each one of the URIs collected was subsequently visited in order to parse and store its content in plaintext format. It should be noted that the system handles multiple document formats available in Web pages (i.e. .html, .doc, php, ppt, pdf, flash etc). After the translation and keyword extraction process, the top-10 keywords assigned to each URI along with their respective frequencies were included in a separate table of the database. The size of the resulting SQL Server database was 891 MB. Additionally, the final keywords’ set along with the keywords extracted using the 3 different methods (content, inlinks, outlinks) before the translation process were also stored in XML files named after the respective URI. In the majority of the Web pages, the results are more than satisfying. An example of the output of this phase is included in the Appendix.

As far as document clustering is concerned, we have performed a set of preliminary experiments. The experiments involved applying both algorithms (K-Means and DBSCAN) with different values of the input parameters to data sets of various length, in a PC Celeron 2.0 GHz, 456MB RAM. Due to space constraints, in this paper we present the results of clustering a data set containing 10.000 documents. The performance, as well as the values of the validity coefficients for the K-means and DBSCAN algorithms, are listed in Table 1 and 2 respectively.

Observing the experimental results, we conclude that for the dataset of the 10000 records DBSCAN was less time consuming than K-means. The modified Hubert Γ statistics achieved its largest value in the K-means algorithm, whereas for DBSCAN the respective values are rather small, meaning that for large datasets K-means creates more compact clusters. The other two indexes, however, prove DBSCAN algorithm advantage, since it creates more separated (distant) and less similar clusters.

Table 1. K-means evaluation

Criteria/Parameters	K-means				
	K - number of clusters				
	5	10	25	50	100
Preprocessing time (msec)	327201	327201	327201	327201	327201
Algorithm execution time (msec)	1272	1132	932	1052	1312
Modified Hubert Statistics	1.545	1.955	2.178	2.820	3.986
Dunn index	0.173	0.157	0.159	0.141	0.149
David-Bouldin index	6.555	6.116	5.388	3.679	1.945

Table 2. DBSCAN evaluation

Criteria /Parameters	DBSCAN					
	ϵ - distance between points, that form cluster/ MinPts – minimal number of points in cluster					
	1.0/4	1.0/2	0.9/3	0.8/2	0.8/3	0.7/2
Preprocessing time (msec)	297708	297708	297708	297708	297708	297708
Algorithm execution time (msec)	481	431	440	430	441	441
Modified Hubert Statistics	0.403	0.697	0.264	0.264	0.138	0.232
Dunn index	0.460	0.408	0.965	1.249	3.984	1.030
David-Bouldin index	2.234	2.295	1.240	0.643	0.167	0.909

The size of the data sets is relatively small compared to the amount of documents residing on the Archive. We should point out, however, that this is a set of preliminary experiments we have performed in order to evaluate the quality of the clusters created when using each algorithm. The next step, as already mentioned, will be to use a hybrid method, which first forms clusters from a part of the documents, and then classifies the rest of them. We also intend to examine the quality of the resulting clusters in terms of semantic coherence.

6. Future Work

The first phase of the Greek Web archiving project has concluded. Most of the Greek URIs have been collected and processed in order to be semantically characterized by a set of keywords. We have already performed a set of preliminary experiments in order

to decide which clustering schema performs better with our data set. We aim at applying more advanced clustering techniques, such as hybrid clustering/classification techniques, in order to achieve the categorization of all the documents that reside on the Greek Web Archive. We also intend to perform a set of blind tests in order to evaluate the quality of the resulting clusters using blind-testing techniques.

We aim at exploiting the characterization of the documents with keywords, and try alternative clustering techniques, utilizing the *semantic proximity* of the documents. Based on the semantic characterization of Web pages, a “search engine” of the Web Archive will also be created. Apart from that, we intend on implementing an automatic crawler (robot), which will periodically visit these Web pages in order to keep their versions as they change through time. We also intend to examine more thoroughly the invisible Web issue.

7. Acknowledgements

We would like to thank S. Maridakis, P. Panousakis and K. Paridis for their help in the crawling process. We would also like to thank S. Pavlakis for his help during the Greek words’ stemming process.

8. References

1. S. Abiteboul, G. Cobena, J. Masanes, G. Sedrati: A First Experience in Archiving the French Web, in Proceedings of the Research and advanced technology for digital libraries: 6th European conference (ECDL 2002), Italy (2002)
2. A. Arvidson, K. Persson, J. Mannerheim: The Kulturarw3 Project – the Royal Swedish Web Archiw3e: an example of ‘complete’ collection of web pages, 66th IFLA Council and General Conference, Israel (2000)
3. S. Brin, L. Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine, in Proceedings of WWW7 (1998)
4. W. Cathro, C. Webb, J. Whiting: Archiving the Web: the PANDORA archive and the National Library of Australia, in Proceedings of the Preserving the Present for the Future Web Archiving Conference, Copenhagen (2001)
5. S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins: Mining the Link Structure of the World Wide Web, IEEE Computer (1999), 32(6)
6. S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, J. Kleinberg: Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text, in Proceedings of WWW7 (1998)
7. M. Eirinaki, M. Vazirgiannis, I. Varlamis: SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process, in Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2003), August 2003, Washington DC
8. M. Eirinaki, C. Lampos, S. Paulakis, M. Vazirgiannis: Web Personalization Integrating Content Semantics and Navigational Patterns, submitted for revision at WIDM’04
9. M. Ester, H.P. Kriegel, J. Sander, M. Wimmer and X. Xu: Incremental Clustering for Mining in a Data Warehousing Environment, in Proceedings of the 24th VLDB Conference (1998)

10. J. Hakala: Collecting and preserving the Web: Developing and Testing the NEDLIB Harvester, RLG DigiNews, 5(2), (2001)
11. M. Halkidi, B. Nguyen, I. Varlamis, M. Vazirgiannis: THESUS: Organizing Web Document Collections Based on Link Semantics, in VLDB Journal, special issue on Semantic Web (2003)
12. M. Halkidi, Y. Batistakis, M. Vazirgiannis: Cluster Validity Methods: Part II, SIGMOD Record, September 2002.
13. Internet Archive, Wayback Machine, <http://www.archive.org>
14. L. Martin: Networked Electronic Publications Policy and Guidelines, Electronic Collections Coordinating Group, National Library of Canada (1998), <http://www.nlc-bnc.ca/9/8/index-e.html>
15. S. Paulakis, H. Lamos, M. Eirinaki, M. Vazirgiannis: SEWeP: A Web Mining System supporting Semantic Personalization, to appear in ECML/PKDD 2004 Proceedings (demo session)
16. The Web archive bibliography, <http://www.ifs.tuwien.ac.at/~aola/links/WebArchiving.html>
17. M. Triantafyllidis, Triantafyllidis On-Line, Modern Greek Language Dictionary, <http://kastor.komvos.edu.gr/dictionaries/dictonline/DictOnLineTri.htm>
18. WordNet, A lexical database for the English language, <http://www.cogsci.princeton.edu/~wn/>
19. Z. Wu, M. Palmer: Verb Semantics and Lexical Selection, 32nd Annual Meetings of the Associations for Computational Linguistics (1994), 133-138

APPENDIX

In this section we include the keywords extracted from four Greek Web pages, covering diverse subjects. Their content is either in Greek, or in English. When Greek content was encountered, we also include the Greek keywords extracted:

Fig. 3. The official site of the Greek Cultural Olympiad (2001-2004)
http://www.cultural-olympiad.gr/1/11_en.html

The screenshot shows an XML viewer interface. The main window displays a tree structure under the root element 'keywords'. The tree is expanded to show the following elements: 'content', 'outlinks', 'total', and 'total_translated'. The 'total_translated' element is further expanded to show a table of 10 keywords and their frequencies.

	word	freq
1	cultural	14
2	event	11
3	olympiad	8
4	olympic	7
5	programme	6
6	game	5
7	athens	4
8	world	3
9	seek	3
10	international	3

Fig. 4. The index page of the National Centre for Scientific Research “Demokritos” http://www.demokritos.gr/index_gr.html

XML		
▲ keywords		
▼ content		
▼ outlinks		
▲ total		
▲ keyword (10)		
word	freq	
1	ινστιτουτο	16
2	πυρηνικός	4
3	πληροφορική	2
4	ελληνικός	2
5	ατομικός	2
6	υλικό	2
7	γραφείο	2
8	βιβλιοθήκη	2
9	ραδιοϊσότοπο	2
10	κέντρο	2
▲ total_translated		
▲ keyword (10)		
word	freq	
1	institute	16
2	nuclear	4
3	library	2
4	stiffening	2
5	informatics	2
6	individual	2
7	office	2
8	radioisotope	2
9	Greek	2
10	centre	2